Linearly Convergent Evolution Strategies via Augmented Lagrangian Constraint Handling

Asma Atamna, Anne Auger, Nikolaus Hansen

Linearly Convergent Evolution Strategies via Augmented Lagrangian Constraint Handling

Asma Atamna
Inria
Centre Saclay–Île-de-France
LRI, Université Paris-Saclay
atamna@lri.fr

Anne Auger
Inria
Centre Saclay–Île-de-France
LRI, Université Paris-Saclay
auger@lri.fr

Nikolaus Hansen
Inria
Centre Saclay–Île-de-France
LRI, Université Paris-Saclay
hansen@lri.fr

ABSTRACT

We analyze linear convergence of an evolution strategy for constrained optimization with an augmented Lagrangian constraint handling approach. We study the case of multiple active linear constraints and use a Markov chain approach—used to analyze randomized optimization algorithms in the unconstrained case—to establish linear convergence under sufficient conditions. More specifically, we exhibit a class of functions on which a homogeneous Markov chain (defined from the state variables of the algorithm) exists and whose stability implies linear convergence. This class of functions is defined such that the augmented Lagrangian, centered in its value at the optimum and the associated Lagrange multipliers, is positive homogeneous of degree 2, and includes convex quadratic functions. Simulations of the Markov chain are conducted on linearly constrained sphere and ellipsoid functions to validate numerically the stability of the constructed Markov chain.

Keywords

Augmented Lagrangian, constrained optimization, evolution strategies, Markov chain, randomized optimization algorithms

1. INTRODUCTION

Randomized (or stochastic) optimization algorithms are robust methods widely used in industry for solving continuous real-world problems. Among them, the covariance matrix adaptation (CMA) evolution strategy (ES) [12] is nowadays recognized as the state-of-the art method. It exhibits linear convergence on wide classes of functions when solving unconstrained optimization problems. However, many practical problems come with constraints and the question of how to handle them properly to particularly preserve the linear convergence is an important one [2]. Recently, an augmented Lagrangian approach to handle constraints within ES algorithms was proposed with the motivation to design an algorithm converging linearly [2]. The algorithm was analyzed theoretically and sufficient conditions for linear convergence, posed in terms of stability conditions of an underlying Markov chain, were formulated [3]. In those works, however, only the case of a single linear constraint was considered.

FOGA '17, January 12 - 15, 2017, Copenhagen, Denmark

ISBN 978-1-4503-4651-1/17/01…$15.00

DOI: http://dx.doi.org/10.1145/3040718.3040732

Markov chain theory [14] provides useful tools to analyze the linear convergence of adaptive randomized optimization algorithms and particularly evolution strategies. In a nutshell, for the case of unconstrained optimization, on scaling-invariant functions—a class of functions that includes all convex-quadratic functions—for adaptive ESs satisfying certain invariance properties (typically translation and scale-invariance), the stability analysis of an appropriate Markov chain can lead to linear convergence proofs of the original algorithm [7]. This general approach was exploited in [5] to prove the linear convergence of the $(1, \lambda)$-ES with self-adaptation on the sphere function and in [6] to prove the linear convergence of the $(1 + 1)$-ES with $1/5$th success rule. This general methodology to prove linear convergence in the case of unconstrained optimization was generalized to constrained optimization, in the case where a single constraint is handled via an adaptive augmented Lagrangian approach [3]. The underlying algorithm being a $(1 + 1)$-ES.

In this work, we generalize the study in [3] to the case of multiple linear inequality constraints. We analyze a $(\mu/\mu_{\mathrm{w}}, \lambda)$-ES with an augmented Lagrangian constraint handling approach in the case of active constraints. The analyzed algorithm is an extension of the one analyzed in [3], where we generalize the original update rule for the penalty factor in [2] to the case of multiple constraints. We construct a homogeneous Markov chain for problems such that the corresponding augmented Lagrangian, centered at the optimum of the problem and the corresponding Lagrange multipliers, is positive homogeneous of degree 2, given some invariance properties are satisfied by the algorithm. Then, we give sufficient stability conditions on the Markov chain such that the algorithm converges to the optimum of the constrained problem as well as to the associated Lagrange multipliers. Finally, the stability of the constructed Markov chain is investigated empirically.

The rest of this paper is organized as it follows: we present augmented Lagrangian methods in Section 2 and give an overview on how the Markov chain approach is used to prove linear convergence in the unconstrained case in Section 3. We formally define the studied optimization problem, as well as the considered augmented Lagrangian in Sections 4 and 5 respectively. In Section 6, we present the studied algorithm and discuss its invariance properties. In Section 7, we present the constructed Markov chain and deduce linear convergence given its stability. Finally, we present our empirical results in Section 8 and conclude with a discussion in Section 9.

Notations

The notations that are not explicitly defined in the paper are presented here. We denote $\mathbb{R}^+$ the set of positive real numbers, $\mathbb{R}^+_>$ the set of strictly positive real numbers, and $\mathbb{N}_>$ the set of natural

numbers without 0. $\mathbf{x} \in \mathbb{R}^n$ is a column vector, $\mathbf{x}^\intercal$ is its transpose, and $\mathbf{0} \in \mathbb{R}^n$ is the zero vector. $\|\mathbf{x}\|$ denotes the Euclidean norm of $\mathbf{x}$, $[\mathbf{x}]_i$ its ith element, and $[\mathbf{M}]_{ij}$ the element in the ith row and jth column of matrix $\mathbf{M}$. $\mathbf{I}_{n\times n} \in R^{n\times n}$ denotes the identity matrix, $\mathcal{N}(\mathbf{0}, \mathbf{I}_{n\times n})$ the multivariate standard normal distribution, and $\sim$ the equality in distribution. The symbol $\circ$ is the function composition operator. The derivative with respect to $\mathbf{x}$ is denoted $\nabla_\mathbf{x}$ and the expectation of a random variable $X \sim \pi$ is denoted E_π.

2. AUGMENTED LAGRANGIAN METHODS

Augmented Lagrangian methods are constraint handling approaches that combine penalty function methods with the Karush-Kuhn-Tucker (KKT) necessary conditions of optimality. They were first introduced in [13, 16] to overcome the limitations of penalty function methods—in particular quadratic penalty methods—which suffer from ill-conditioning as the penalty parameters need to tend to infinity in order to converge [15].

Similarly to penalty methods, augmented Lagrangian methods transform the constrained problem into one or more unconstrained problems where an augmented Lagrangian, consisting in a Lagrangian part and a penalty function part, is optimized. The Lagrangian is a function $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ defined as

$$\mathcal{L}(\mathbf{x}, \gamma) = f(\mathbf{x}) + \sum_{i=1}^{m} \gamma^i g_i(\mathbf{x}) \ , \tag{1}$$

for a function f subject to m constraints $g_i(\mathbf{x}) \leq 0$. The vector $\gamma = (\gamma^1, \cdots, \gamma^m)^\intercal \in \mathbb{R}^m$ represents the Lagrange factors. An important property of $\mathcal{L}$ is the so-called KKT stationarity condition which states that, given some regularity conditions (constraint qualifications) are satisfied, if $\mathbf{x}^* \in \mathbb{R}^n$ is a local optimum of the constrained problem, then there exists a vector $\gamma^* = (\gamma^{*1}, \cdots, \gamma^{*m})^\intercal \in (\mathbb{R}^+)^m$ of Lagrange multipliers γ^{*i}, $i = 1, \cdots, m$, such that

$$\nabla_\mathbf{x}\mathcal{L}(\mathbf{x}^*, \gamma^*) = \nabla_\mathbf{x} f(\mathbf{x}^*) + \sum_{i=1}^{m} \gamma^{*i} \nabla g_i(\mathbf{x}^*) = \mathbf{0} \ ,$$

if we assume f and g_i, $i = 1, \cdots, m$, are differentiable at $\mathbf{x}^*$.

Remark 1. Given the gradients $\nabla_\mathbf{x} f(\mathbf{x}^*)$ and $\nabla_\mathbf{x} g_i(\mathbf{x}^*)$, $i = 1, \cdots, m$, exist, the first-order necessary conditions of optimality (KKT conditions) ensure the existence of at least one vector γ^* of Lagrange multipliers. However, if the constraints satisfy the linear independence constraint qualification (LICQ), that is, the set of constraint normals is linearly independent, the vector γ^* of Lagrange multipliers is unique [15].

The Lagrangian $\mathcal{L}$ is combined to a penalty function, which is a function of the constraints g_i, to construct the augmented Lagrangian h. Examples of augmented Lagrangians are given in (9) and (10), where $\omega = (\omega^1, \cdots, \omega^m)^\intercal \in (\mathbb{R}^+_>)^m$ is the vector of the penalty factors ω^i. More generally, the augmented Lagrangian can be defined as

$$h(\mathbf{x}, \gamma, \omega) = f(\mathbf{x}) + \sum_{i=1}^{m} \varphi(g_i(\mathbf{x}), \gamma^i, \omega^i) \ , \tag{2}$$

where φ is chosen such that a local optimum $\mathbf{x}^*$ of the constrained problem is a stationary point of h, that is for all $\omega \in (\mathbb{R}^+_>)^m$,

$$\nabla_\mathbf{x} h(\mathbf{x}^*, \gamma^*, \omega) = \mathbf{0} \ ,$$

assuming the gradients at $\mathbf{x}^*$ are defined. The augmented Lagrangian h is minimized for given values of γ and ω instead of the initial objective function f.

In adaptive augmented Lagrangian approaches, γ is adapted to approach the Lagrange multipliers and ω is adapted to guide the search towards feasible solutions. A proper adaptation mechanism for ω helps preventing ill-conditioning since, with an augmented Lagrangian approach, the penalty factors ω^i do not need to tend to infinity to achieve convergence [15].

There exist in the literature some examples where augmented Lagrangian approaches are used in the context of evolutionary algorithms. In [17], the authors present a coevolutionary method for constrained optimization with an augmented Lagrangian approach, where two populations (one for the parameter vector and one for Lagrange factors) are evolved in parallel, using an evolution strategy with self-adaptation. The approach is tested on four non-linear constrained problems, with a fixed value for the penalty parameter.

In [9], the authors present an augmented-Lagrangian-based genetic algorithm for constrained optimization. Their algorithm requires a local search procedure for improving the current best solution in order to converge to the optimal solution and to the associated Lagrange multipliers.

More recently, an augmented Lagrangian approach was combined with a $(1 + 1)$-ES for the case of a single linear constraint [2]. An update rule was presented for the penalty parameter and the algorithm was observed to converge on the sphere function and on a moderately ill-condition ellipsoid function, with one linear constraint. This algorithm was analyzed in [3] using tools from the Markov chain theory. The authors constructed a homogeneous Markov chain and deduced linear convergence under the stability of this Markov chain. In [4], the augmented Lagrangian constraint handling mechanism in [2] is implemented for CMA-ES and a general framework for building a general augmented Lagrangian based randomized algorithm for constrained optimization in the case of one constraint is presented.

3. MARKOV CHAIN ANALYSIS AND LINEAR CONVERGENCE

Randomized or stochastic optimization algorithms are iterative methods where—most often—the state of the algorithm is a Markov chain. For a certain class of algorithms obeying proper invariance properties, Markov chain theory can provide powerful tools to prove the linear convergence of the algorithms [8, 7, 5]. We illustrate here on a simple case the general methodology to prove linear convergence of an adaptive randomized algorithm using Markov chain theory. We assume for the sake of simplicity the minimization of the sphere function $\mathbf{x} \mapsto f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\intercal\mathbf{x}$ with, without loss of generality, the optimum in zero. We assume that the state of the algorithm at iteration t is given by the current estimate $\mathbf{X}_t$ of the optimum and a positive factor, the step-size σ_t. From this state, λ new candidate solutions are sampled according to

$$\mathbf{X}^i_{t+1} = \mathbf{X}_t + \sigma_t \mathbf{U}^i_{t+1}, \ i = 1, \ldots, \lambda \ ,$$

where $\mathbf{U}^i_{t+1}$ are independent identically distributed (i.i.d.) standard multivariate normal distributions (with mean zero and covariance matrix identity). The state of the algorithm is then updated via two deterministic update functions $\mathcal{G}_\mathbf{x}$ and $\mathcal{G}_\sigma$ according to

$$\mathbf{X}_{t+1} = \mathcal{G}_\mathbf{x}((\mathbf{X}_t, \sigma_t), \varsigma * \mathbf{U}_{t+1}) \ , \tag{3}$$

$$\sigma_{t+1} = \mathcal{G}_\sigma(\sigma_t, \varsigma * \mathbf{U}_{t+1}) \ , \tag{4}$$

where $\mathbf{U}_{t+1} = [\mathbf{U}^1_{t+1}, \cdots, \mathbf{U}^\lambda_{t+1}]$ is the vector of i.i.d. random vectors $\mathbf{U}^i_{t+1}$ and

$$\varsigma = Ord(f(\mathbf{X}_t + \sigma_t \mathbf{U}^i_{t+1})_{i=1,\cdots,\lambda})$$

is the permutation that contains the indices of the candidate solutions $\mathbf{X}_t + \sigma_t \mathbf{U}_{t+1}^i$ ranked according to their f-value. That is, the ordering is done using the operator *Ord* such that, given λ real numbers $z_1, \cdots, z_\lambda$, $\varsigma = Ord(z_1, \cdots, z_\lambda)$ satisfies

$$z_{\varsigma(1)} \leq \cdots \leq z_{\varsigma(\lambda)} \ . \tag{5}$$

In (3) and (4), the operator '*' applies the permutation ς to $\mathbf{U}_{t+1}$ and

$$\varsigma * \mathbf{U}_{t+1} = [\mathbf{U}_{t+1}^{\varsigma(1)}, \cdots, \mathbf{U}_{t+1}^{\varsigma(\lambda)}] \ . \tag{6}$$

It has been shown that if the update functions $\mathcal{G}_\mathbf{x}$ and $\mathcal{G}_\sigma$ satisfy the following conditions [7]:

(i) for all $\mathbf{x}, \mathbf{x}_0 \in \mathbb{R}^n$, for all $\sigma > 0$, for all $\mathbf{y} \in (\mathbb{R}^n)^\lambda$

$$\mathcal{G}_\mathbf{x}((\mathbf{x} + \mathbf{x}_0, \sigma), \mathbf{y}) = \mathcal{G}_\mathbf{x}((\mathbf{x}, \sigma), \mathbf{y}) + \mathbf{x}_0 \ ,$$

(ii) for all $\mathbf{x} \in \mathbb{R}^n$, for all $\alpha, \sigma > 0$, for all $\mathbf{y} \in (\mathbb{R}^n)^\lambda$

$$\mathcal{G}_\mathbf{x}((\mathbf{x}, \sigma), \mathbf{y}) = \alpha \mathcal{G}_\mathbf{x}\Big(\Big(\frac{\mathbf{x}}{\alpha}, \frac{\sigma}{\alpha}\Big), \mathbf{y}\Big) \ ,$$

(iii) for all $\alpha, \sigma > 0$, for all $\mathbf{y} \in (\mathbb{R}^n)^\lambda$

$$\mathcal{G}_\sigma(\sigma, \mathbf{y}) = \alpha \mathcal{G}_\sigma\Big(\frac{\sigma}{\alpha}, \mathbf{y}\Big) \ ,$$

then the algorithm is translation-invariant and scale-invariant. As a consequence, $(\mathbf{Y}_t)_{t\in\mathbb{N}}$, with $\mathbf{Y}_t = \frac{\mathbf{X}_t}{\sigma_t}$, is a homogeneous Markov chain that can be defined independently of $(\mathbf{X}_t, \sigma_t)$, given $\mathbf{Y}_0 = \frac{\mathbf{X}_0}{\sigma_0}$, as

$$\mathbf{Y}_{t+1} = \frac{\mathcal{G}_\mathbf{x}((\mathbf{Y}_t, 1), \varsigma * \mathbf{U}_{t+1})}{\mathcal{G}_\sigma(1, \varsigma * \mathbf{U}_{t+1})} \ ,$$

where $\varsigma = Ord(f(\mathbf{Y}_t + \mathbf{U}_{t+1}^i)_{i=1,\cdots,\lambda})$ [7, Proposition 4.1] (this result is true for the sphere function but more generally for a scaling-invariant objective function). Let consider now the following definition of linear convergence:

Definition 1. We say that a sequence $(\mathbf{X}_t)_{t\in\mathbb{N}}$ of random vectors $\mathbf{X}_t$ converges linearly almost surely (a.s.) to $\mathbf{x}_{\text{opt}}$ if there exists $\text{CR} > 0$ such that

$$\lim_{t\to\infty} \frac{1}{t} \ln \frac{\|\mathbf{X}_t - \mathbf{x}_{\text{opt}}\|}{\|\mathbf{X}_0 - \mathbf{x}_{\text{opt}}\|} = -\text{CR} \text{ a.s.}$$

The constant CR is called the convergence rate.

Using the property of the logarithm, the quantity $\frac{1}{t} \ln \frac{\|\mathbf{X}_t\|}{\|\mathbf{X}_0\|}$ ($\mathbf{x}_{\text{opt}} = \mathbf{0}$ here) can be expressed as a function of $\mathbf{Y}_t$ according to

$$\begin{aligned}
\frac{1}{t} \ln \frac{\|\mathbf{X}_t\|}{\|\mathbf{X}_0\|} &= \frac{1}{t} \sum_{k=0}^{t-1} \ln \frac{\|\mathbf{X}_{k+1}\|}{\|\mathbf{X}_k\|} \\
&= \frac{1}{t} \sum_{k=0}^{t-1} \ln \frac{\|\mathbf{X}_{k+1}\|}{\|\mathbf{X}_k\|} \frac{\sigma_k \mathcal{G}_\sigma(1, \varsigma * \mathbf{U}_{k+1})}{\sigma_{k+1}} \\
&= \frac{1}{t} \sum_{k=0}^{t-1} \ln \frac{\|\mathbf{Y}_{k+1}\|}{\|\mathbf{Y}_k\|} \mathcal{G}_\sigma(1, \varsigma * \mathbf{U}_{k+1}) \ ,
\end{aligned} \tag{7}$$

where we have successively artificially introduced $\sigma_{k+1} = \sigma_k \mathcal{G}_\sigma(1, \varsigma * \mathbf{U}_{k+1})$ and then used that $\mathbf{Y}_k = \mathbf{X}_k/\sigma_k$ and $\mathbf{Y}_{k+1} = \mathbf{X}_{k+1}/\sigma_{k+1}$. In (7), we have expressed the term whose limit we are interested in as the empirical average of a function of a Markov chain. However, we know from Markov chain theory that if some sufficient stability conditions—given for instance in Theorem 17.0.1 from [14]—are satisfied by $(\mathbf{Y}_t)_{t\in\mathbb{N}}$, then a law of large numbers (LLN) for Markov chains can be applied to the right-hand side of the previous equation. Consequently,

$$\begin{aligned}
\lim_{t\to\infty} \frac{1}{t} \ln \frac{\|\mathbf{X}_t\|}{\|\mathbf{X}_0\|} &= \lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} \ln \frac{\|\mathbf{Y}_{k+1}\|}{\|\mathbf{Y}_k\|} \mathcal{G}_\sigma(1, \varsigma * \mathbf{U}_{k+1}) \\
&= \int \ln \|\mathbf{y}\| \pi(d\mathbf{y}) - \int \ln \|\mathbf{y}\| \pi(d\mathbf{y}) \\
&\quad + \underbrace{\int E(\ln(\mathcal{G}_\sigma(1, \varsigma * \mathbf{U}_{t+1})) | \mathbf{Y}_t = \mathbf{y}) \pi(d\mathbf{y})}_{-\text{CR}} \ ,
\end{aligned}$$

where π is the invariant probability measure of the Markov chain $(\mathbf{Y}_t)_{t\in\mathbb{N}}$. Hence, assuming that a law of large number holds for the Markov chain $(\mathbf{Y}_t)_{t\in\mathbb{N}}$, the algorithm described by the iterative sequence $(\mathbf{X}_t, \sigma_t)_{t\in\mathbb{N}}$ will converge linearly at the rate expressed as minus the expected log step-size change (where the expectation is taken with respect to the invariant probability measure of $(\mathbf{Y}_t)_{t\in\mathbb{N}}$). This methodology to prove the linear convergence of adaptive algorithms (including many evolution strategies) in the unconstrained case holds on scaling-invariant functions (that include particularly functions that write $g \circ f$, where g is a 1-D strictly increasing function and f is positively homogeneous, typically f can be a convex-quadratic function). It provides the *exact* expression of the convergence rate that equals the expected log step-size change with respect to the stationary distribution of a Markov chain. This illustrates that Markov chains are central tools for the analysis of convergence of adaptive randomized optimization algorithms. Remark that the convergence rate can be easily simulated to obtain quantitative estimates and dependencies with respect to internal parameters of the algorithm or of the objective functions.

We see that there are two distinct steps for the analysis of the linear convergence:

(i) Identify on which class of functions the algorithms we study can exhibit a Markov chain whose stability will lead to the linear convergence of the underlying algorithm (in the example above, the Markov chain equals $\mathbf{Y}_t = \mathbf{X}_t/\sigma_t$).

(ii) Prove the stability of the identified Markov chain.

The second step is arguably the most complex one. So far, it has been successfully achieved for the analysis of the linear convergence of self-adaptive evolution strategies [5] and for the $(1+1)$-ES with one-fifth success rule [6] in the unconstrained case. The main tools to prove the stability rely on Foster-Lyapunov drift conditions [14]. In this paper, we will focus on the first step. Particularly, the Markov chain for step-size adaptive randomized search optimizing scaling-invariant functions (i.e. unconstrained optimization) was identified in [7]. In addition, in the constrained case, the Markov chain has been identified for the $(1 + 1)$-ES with an augmented Lagrangian constraint handling in the case of one linear inequality constraint [3]. We consider here the extension to more than one constraint and a more general algorithm framework.

4. OPTIMIZATION PROBLEM

We consider throughout this work the problem of minimizing a function f subject to m linear inequality constraints $g_i(\mathbf{x}) \leq 0$, $i = 1, \cdots, m$. Formally, this writes

$$\begin{aligned}
&\min_\mathbf{x} f(\mathbf{x}) \\
&\text{subject to } g_i(\mathbf{x}) \leq 0, \ i = 1, \cdots, m \ ,
\end{aligned} \tag{8}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $g_i : \mathbb{R}^n \to \mathbb{R}$, and $g_i(\mathbf{x}) = \mathbf{b}_i{}^\intercal \mathbf{x} + c_i$, $\mathbf{b}_i \in \mathbb{R}^n$, $c_i \in \mathbb{R}$.

We assume this problem to have a unique global optimum $\mathbf{x}_{\text{opt}}$. We also assume the constraints to be active at $\mathbf{x}_{\text{opt}}$, that is, $g_i(\mathbf{x}_{\text{opt}}) = 0$, $i = 1, \cdots, m$. This constitutes the most difficult case. Indeed, if the constraint is not active, when close enough to the optimum, the algorithm will typically not see the constraint such that it will behave as in the unconstrained case. In terms of theoretical analysis, the unconstrained case—for a general class of step-size adaptive algorithms—is well understood in the case of scaling-invariant functions [7]. Additionally, we assume that the gradients at $\mathbf{x}_{\text{opt}}$, $\nabla_{\mathbf{x}} f(\mathbf{x}_{\text{opt}})$ and $\nabla_{\mathbf{x}} g_i(\mathbf{x}_{\text{opt}})$, $i = 1, \cdots, m$, are defined and that the constraints satisfy the linear independence constraint qualification (LICQ) [15] at $\mathbf{x}_{\text{opt}}$. We denote γ_{opt} the (unique) vector of Lagrange multipliers associated to $\mathbf{x}_{\text{opt}}$.

5. CONSIDERED AUGMENTED LAGRANGIAN

A practical augmented Lagrangian for the optimization problem in (8) is given in the following equation

$$h(\mathbf{x}, \gamma, \omega) = f(\mathbf{x}) + \sum_{i=1}^{m} \underbrace{\begin{cases} \gamma^i g_i(\mathbf{x}) + \frac{\omega^i}{2} g_i(\mathbf{x})^2 & \text{if } \gamma^i + \omega^i g_i(\mathbf{x}) \geq 0 \\ -\frac{{\gamma^i}^2}{2\omega^i} & \text{otherwise} \end{cases}}_{\varphi_1(g_i(\mathbf{x}), \gamma^i, \omega^i)} \quad . \tag{9}$$

The use of a different penalty factor for each constraint is motivated by the fact that the penalization should depend on the constraint violation—which might be different for different constraints. The quality of a solution $\mathbf{x}$ is evaluated by adding $f(\mathbf{x})$ and either (i) $\gamma^i g_i(\mathbf{x}) + \frac{\omega^i}{2} g_i(\mathbf{x})^2$ if $g_i(\mathbf{x})$ is larger than $-\frac{\gamma^i}{\omega^i}$ or (ii) $-\frac{{\gamma^i}^2}{2\omega^i}$ otherwise, for each constraint function g_i.

The augmented Lagrangian in (9) is constructed such that (i) the fitness function remains unchanged when far in the feasible domain and (ii) h is "smooth" in that it is differentiable with respect to g_i. Therefore, (9) is the recommended augmented Lagrangian in practice. For the analysis, however, we consider a simpler augmented Lagrangian (equation below) so that we can construct a Markov chain.

$$h(\mathbf{x}, \gamma, \omega) = f(\mathbf{x}) + \sum_{i=1}^{m} \underbrace{\gamma^i g_i(\mathbf{x}) + \frac{\omega^i}{2} g_i(\mathbf{x})^2}_{\varphi_2(g_i(\mathbf{x}), \gamma^i, \omega^i)} \quad . \tag{10}$$

The difference is that in the previous formulation the penalization is a constant and hence inconsequential for $g_i(\mathbf{x}) < -\gamma^i/\omega^i$. Since we focus in our study on problems where the constraints are active at the optimum, the augmented Lagrangians in (9) and (10) are equivalent in the vicinity of $\mathbf{x}_{\text{opt}}$, as illustrated in Figure 1 for one constraint. Inactive constraints are covered in that the analysis remains valid when these constraints are removed, in which case we recover the original equation (9) up to adding a constant to the f-value. Therefore, conducting the analysis with (10) gives insight into how a practical algorithm using (9) would perform close to the optimum.

6. ALGORITHM

In this section, we present a general ES (Algorithm 1) with comma-selection and weighted recombination (denoted $(\mu/\mu_{\text{w}}, \lambda)$-ES) for constrained optimization, where the constraints are handled using an augmented Lagrangian approach.

First, λ i.i.d. vectors $\mathbf{U}_{t+1}^i$ are sampled in Line 3 of Algorithm 1 according to a normal distribution of mean $\mathbf{0}$ and covariance matrix

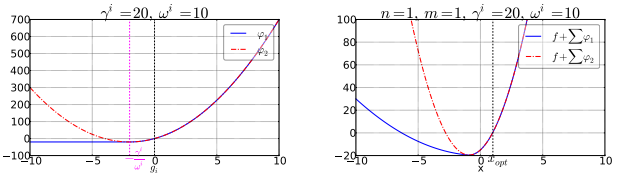

Figure 1: Left: $\varphi_j(g_i(\mathbf{x}), \gamma^i, \omega^i)$ for $j = 1$ (blue) and $j = 2$ (red), as a function of g_i. Right: Augmented Lagrangians, $f(\mathbf{x}) + \sum_{i=1}^{m} \varphi_j(g_i(\mathbf{x}), \gamma^i, \omega^i)$, for $j = 1$ (blue) and $j = 2$ (red), in $n = 1$ with $m = 1$. $f(x) = \frac{1}{2}x^2$, $g_1(x) = x - 1$, and $x_{\text{opt}} = 1$.

the identity. They are used to create λ candidate solutions $\mathbf{X}_{t+1}^i$ according to

$$\mathbf{X}_{t+1}^i = \mathbf{X}_t + \sigma_t \mathbf{U}_{t+1}^i \quad , \tag{11}$$

where $\mathbf{X}_t$ is the current estimate of the optimum and σ_t is the step-size. The candidate solutions are then ranked according to their fitness, determined by their h-value. This is done in Line 4 with the operator *Ord* defined in (5), where ς is the permutation that contains the indices of the ordered candidate solutions.

Later on, we will make explicit the dependency of ς on the objective function, the current solution, and the current step-size, where needed (this would read $\varsigma_{(\mathbf{X}_t, \sigma_t)}^{h(\cdot, \gamma_t, \omega_t)}$ here). The solution $\mathbf{X}_{t+1}$ at the next iteration is computed by recombining the μ best candidate solutions—or parents—in a weighted sum according to Line 5, where w_i, $i = 1, \cdots, \mu$, are the weights associated to the parents and the operator '$*$' applies the permutation ς to the vector $\mathbf{U}_{t+1}$ of the sampled vectors $\mathbf{U}_{t+1}^i$ as defined in (6).

The step-size is adapted in Line 6 using a general update function $\mathcal{G}_\sigma$. For the sake of simplicity, we consider that $\mathcal{G}_\sigma$ is a function of the current step-size σ_t and the ranked vector $\varsigma * \mathbf{U}_{t+1}$ of the sampled vectors $\mathbf{U}_{t+1}^i$.

The Lagrange factors are adapted in Line 7. As a result of this update rule, a Lagrange factor γ_t^i is increased if $g_i(\mathbf{X}_{t+1})$ is positive and decreased otherwise. A damping factor d_γ is used to attenuate the change in the value of γ_t^i.

Each penalty factor ω_t^i is adapted according to Line 8. This update is a generalization to the case of many constraints of the original update proposed in [2] for the case of a single constraint. A penalty factor ω_t^i is increased in two cases: the first one is given by the first inequality in Line 8 and corresponds to the case where (i) the change in h-value due to changes in γ_t^i and ω_t^i is smaller than the change in h-value due to the change in $\mathbf{X}_t$. Indeed

$$\omega_t^i g_i(\mathbf{X}_{t+1})^2 \approx |h(\mathbf{X}_{t+1}, \gamma_t + \Delta_i\gamma, \omega_t + \Delta_i\omega) - h(\mathbf{X}_{t+1}, \gamma_t, \omega_t)| \quad ,$$

where $\Delta_i\gamma = (0, \cdots, \Delta\gamma^i, \cdots, 0)^\intercal$ and $\Delta_i\omega = (0, \cdots, \Delta\omega^i, \cdots, 0)^\intercal$. By increasing the penalization, we prevent premature stagnation [2]. The parameter ω_t^i is also increased if (ii) the change in the corresponding constraint value $|g_i(\mathbf{X}_{t+1}) - g_i(\mathbf{X}_t)|$ is significantly smaller than $|g_i(\mathbf{X}_t)|$ (second inequality in Line 8). In this case, increasing the penalization allows approaching the constraint boundary ($g_i(\mathbf{x}) = 0$) more quickly. However, increasing ω_t^i increases the ill-conditioning of the problem at hand, therefore, in all other cases, ω_t^i is decreased (second case in Line 8). Similarly to the update of the Lagrange factors, we use a damping factor d_ω to moderate the changes in ω_t^i.

Algorithm 1 is a randomized adaptive algorithm that can be defined in an abstract manner as follows: given the state variables $(\mathbf{X}_t, \sigma_t, \gamma_t, \omega_t)$ at iteration t, a transition function $\mathcal{F}^{(f, \{g_i\}_{i=1,\cdots,m})}$, and the vector $\mathbf{U}_{t+1} = [\mathbf{U}_{t+1}^1, \cdots, \mathbf{U}_{t+1}^\lambda]$ of i.i.d. normal vectors

Algorithm 1 $(\mu/\mu_{\mathrm{w}}, \lambda)$-ES with Augmented Lagrangian Constraint Handling

0 **given** $n \in \mathbb{N}_>,\ \chi, k_1, k_2, d_\gamma, d_\omega \in \mathbb{R}^+_>, \lambda, \mu \in \mathbb{N}_>, 0 \leq w_i < 1, \sum_{i=1}^{\mu} w_i = 1$

1 **initialize** $\mathbf{X}_0 \in \mathbb{R}^n,\ \sigma_0 \in \mathbb{R}^+_>,\ \gamma_0 \in \mathbb{R}^m,\ \omega_0 \in (\mathbb{R}^+_>)^m,\ t = 0$

2 **while** stopping criterion not met

3 $\quad \mathbf{U}^i_{t+1} = \mathcal{N}(\mathbf{0}, \mathbf{I}_{n\times n}),\ i = 1, \cdots, \lambda$

4 $\quad \varsigma = Ord(h(\mathbf{X}_t + \sigma_t \mathbf{U}^i_{t+1}, \gamma_t, \omega_t)_{i=1,\cdots,\lambda})$

5 $\quad \mathbf{X}_{t+1} = \mathbf{X}_t + \sigma_t \sum_{i=1}^{\mu} w_i [\varsigma * \mathbf{U}_{t+1}]_i, \mathbf{U}_{t+1} = [\mathbf{U}^1_{t+1}, \cdots, \mathbf{U}^\lambda_{t+1}]$

6 $\quad \sigma_{t+1} = \mathcal{G}_\sigma(\sigma_t, \varsigma * \mathbf{U}_{t+1})$

7 $\quad \gamma^i_{t+1} = \gamma^i_t + \frac{\omega^i_t}{d_\gamma} g_i(\mathbf{X}_{t+1}),\ i = 1, \cdots, m$

8 $\quad \omega^i_{t+1} = \begin{cases} \omega^i_t \chi^{1/(4d_\omega)} & \text{if } \omega^i_t g_i(\mathbf{X}_{t+1})^2 < k_1 \times \frac{|h(\mathbf{X}_{t+1},\gamma_t,\omega_t) - h(\mathbf{X}_t,\gamma_t,\omega_t)|}{n} \\ & \text{or } k_2 |g_i(\mathbf{X}_{t+1}) - g_i(\mathbf{X}_t)| < |g_i(\mathbf{X}_t)| \\ \omega^i_t \chi^{-1/d_\omega} & \text{otherwise, } i = 1, \cdots, m \end{cases}$

9 $\quad t = t + 1$

$\mathbf{U}^i_{t+1}$, compute the state at iteration $t+1$ according to

$$(\mathbf{X}_{t+1}, \sigma_{t+1}, \gamma_{t+1}, \omega_{t+1}) = \mathcal{F}^{(f,\{g_i\}_{i=1,\cdots,m})}((\mathbf{X}_t, \sigma_t, \gamma_t, \omega_t), \mathbf{U}_{t+1}),$$

where the superscript indicates the objective function to minimize, f, and the constraint functions, g_i. The deterministic transition function $\mathcal{F}^{(f,\{g_i\}_{i=1,\cdots,m})}$ is defined by the following general update rules for $\mathbf{X}_t$, σ_t, γ_t, and ω_t:

$$\mathbf{X}_{t+1} = \mathcal{G}_{\mathbf{x}}((\mathbf{X}_t, \sigma_t), \varsigma * \mathbf{U}_{t+1}) \ , \tag{12}$$

$$\sigma_{t+1} = \mathcal{G}_\sigma(\sigma_t, \varsigma * \mathbf{U}_{t+1}) \ , \tag{13}$$

$$\gamma^i_{t+1} = \mathcal{H}^{g_i}_\gamma(\gamma^i_t, \omega^i_t, \mathbf{X}_{t+1}),\ i = 1, \cdots, m \ , \tag{14}$$

$$\omega^i_{t+1} = \mathcal{H}^{(f,g_i)}_\omega(\omega^i_t, \gamma^i_t, \mathbf{X}_t, \mathbf{X}_{t+1}),\ i = 1, \cdots, m \ , \tag{15}$$

where ς, $\mathcal{G}_{\mathbf{x}}$, $\mathcal{H}_\gamma$, and $\mathcal{H}_\omega$ are given in Lines 4, 5, 7, and 8 of Algorithm 1 respectively. These notations are particularly useful for defining the notions of translation and scale-invariance in the next subsection. They also make the connection between the constructed homogeneous Markov chain and the original algorithm clearer.

Comparing (12), (13), (14), and (15) to (3) and (4), it is easy to see that Algorithm 1 is built by taking an adaptive algorithm for unconstrained optimization and changing its objective function to an adaptive one—the augmented Lagrangian—where the parameters of the augmented Lagrangian are additionally adapted every iteration. This idea was already put forward in [4] for the case of a single constraint, and we generalize it here to the case of m constraints.

6.1 Invariance

Invariance with respect to transformations of the search space is a central property in randomized adaptive algorithms. In the unconstrained case, it is exploited to demonstrate linear convergence [7, 6]. In this subsection, we discuss translation-invariance and scale-invariance of Algorithm 1. We first recall the definition of a group homomorphism and introduce some notations.

Definition 2. Let $(G_1, .)$ and $(G_2, *)$ be two groups. A function $\Phi : G_1 \to G_2$ is a group homomorphism if for all $x, y \in G_1$, $\Phi(x.y) = \Phi(x) * \Phi(y)$.

We denote $\mathcal{S}(\Omega)$ the set of all bijective transformations from a set Ω to itself and $\mathrm{Homo}((\mathbb{R}^n, +), (\mathcal{S}(\Omega), \circ))$ (respectively $\mathrm{Homo}((\mathbb{R}^+_>, .), (\mathcal{S}(\Omega), \circ))$) the set of group homomorphisms from $(\mathbb{R}^n, +)$ (respectively from $(\mathbb{R}^+_>, .)$) to $(\mathcal{S}(\Omega), \circ)$.

Translation-invariance informally translates the non-sensitivity of an algorithm with respect to the choice of its initial point, that is the algorithm will exhibit the same behavior when optimizing $\mathbf{x} \mapsto f(\mathbf{x})$ or $\mathbf{x} \mapsto f(\mathbf{x} - \mathbf{x}_0)$ for any $\mathbf{x}_0$. More formally, an algorithm is translation-invariant if we can find a state-space transformation such that optimizing $\mathbf{x} \mapsto f(\mathbf{x})$ or $\mathbf{x} \mapsto f(\mathbf{x} - \mathbf{x}_0)$ is the same up to the state-space transformation. In the next definition, which is a generalization to the constrained case of the definition given in [7], we ask that the set of state-space transformations is given via a group homomorphism from the group acting on the function to transform the functions, that is $(\mathbb{R}^n, +)$, to the group of bijective state-space transformations. Indeed this group homomorphism naturally emerges when attempting to prove invariance. More formally, we have the following definition of translation-invariance.

Definition 3. A randomized adaptive algorithm with transition function $\mathcal{F}^{(f,\{g_i\}_{i=1,\cdots,m})} : \Omega \times \mathbb{U}^\lambda \to \Omega$, where f is the objective function to minimize and g_i are the constraint functions, is translation-invariant if there exists a group homomorphism $\Phi \in \mathrm{Homo}((\mathbb{R}^n, +), (\mathcal{S}(\Omega), \circ))$ such that for any objective function f, for any constraint g_i, for any $\mathbf{x}_0 \in \mathbb{R}^n$, for any state $\mathbf{s} \in \Omega$, and for any $\boldsymbol{u} \in \mathbb{U}^\lambda$,

$$\mathcal{F}^{\left(f(\mathbf{x}),\{g_i(\mathbf{x})\}_{i=1,\cdots,m}\right)}(\mathbf{s}, \boldsymbol{u}) = \Phi(-\mathbf{x}_0)\left(\mathcal{F}^{\left(f(\mathbf{x}-\mathbf{x}_0),\{g_i(\mathbf{x}-\mathbf{x}_0)\}_{i=1,\cdots,m}\right)}(\Phi(\mathbf{x}_0)(\mathbf{s}), \boldsymbol{u})\right) \ .$$

Similarly for scale-invariance, the set of state-space transformations comes from a group homomorphism between the group where the scaling factors acting to transform the objective functions are taken from, that is the group $(\mathbb{R}^+_>, .)$ and the group of bijective state-space transformations.

Definition 4. A randomized adaptive algorithm with transition function $\mathcal{F}^{(f,\{g_i\}_{i=1,\cdots,m})} : \Omega \times \mathbb{U}^\lambda \to \Omega$, where f is the objective function to minimize and g_i are the constraint functions, is scale-invariant if there exists a group homomorphism $\Phi \in \mathrm{Homo}((\mathbb{R}^+_>, .), (\mathcal{S}(\Omega), \circ))$ such that for any objective function f, for any constraint g_i, for any $\alpha > 0$, for any state $\mathbf{s} \in \Omega$, and for any $\boldsymbol{u} \in \mathbb{U}^\lambda$,

$$\mathcal{F}^{\left(f(\mathbf{x}),\{g_i(\mathbf{x})\}_{i=1,\cdots,m}\right)}(\mathbf{s}, \boldsymbol{u}) = \Phi(1/\alpha)\left(\mathcal{F}^{\left(f(\alpha\mathbf{x}),\{g_i(\alpha\mathbf{x})\}_{i=1,\cdots,m}\right)}(\Phi(\alpha)(\mathbf{s}), \boldsymbol{u})\right) \ .$$

The next proposition states translation-invariance of Algorithm 1.

PROPOSITION 1. *Algorithm 1 is translation-invariant and the associated group homomorphism Φ is given by*

$$\Phi(\boldsymbol{x}_0)(\boldsymbol{x}, \sigma, \gamma, \omega) = (\boldsymbol{x} + \boldsymbol{x}_0, \sigma, \gamma, \omega) \ , \tag{16}$$

for all $\boldsymbol{x}_0, \boldsymbol{x} \in \mathbb{R}^n$, for all $\sigma \in \mathbb{R}$, and for all $\gamma, \omega \in \mathbb{R}^m$.

The proof of this proposition is given in Appendix A.1. In the next proposition we state the scale-invariance of Algorithm 1 under scale-invariance of the transition function $\mathcal{G}_\sigma$.

PROPOSITION 2. *If the update function $\mathcal{G}_\sigma$ of the step-size satisfies the following condition*

$$\mathcal{G}_\sigma(\sigma_t, \varsigma * \boldsymbol{U}_{t+1}) = \alpha \mathcal{G}_\sigma(\sigma_t/\alpha, \varsigma * \boldsymbol{U}_{t+1}) \ , \tag{17}$$

for all $\alpha > 0$, then Algorithm 1 is scale-invariant and the associated group homomorphism Φ is defined as

$$\Phi(\alpha)(\boldsymbol{x}, \sigma, \gamma, \omega) = (\boldsymbol{x}/\alpha, \sigma/\alpha, \gamma, \omega) \ , \tag{18}$$

for all $\alpha > 0$, for all $\boldsymbol{x} \in \mathbb{R}^n$, for all $\sigma \in \mathbb{R}$, and for all $\gamma, \omega \in \mathbb{R}^m$.

The proof of the proposition is given in Appendix A.2.

In the next section, we illustrate how translation and scale-invariance induce the existence of a homogeneous Markov chain whose stability implies linear convergence.

7. ANALYSIS

In this section, we demonstrate the existence of an underlying homogeneous Markov chain to Algorithm 1, given the augmented Lagrangian in (10) satisfies a particular condition. To construct the Markov chain, we exploit invariance properties of Algorithm 1, as well as the updates of the Lagrange factors and the penalty factors.

As stated in Section 4, we assume that the optimization problem admits a unique global optimum $\mathbf{x}_{\text{opt}}$ and that the constraints g_i, $i = 1, \cdots, m$, satisfy the LICQ at $\mathbf{x}_{\text{opt}}$, hence that the vector γ_{opt} of Lagrange multipliers is unique. Once we have the Markov chain, we show how its stability leads to linear convergence of (i) the current solution $\mathbf{X}_t$ towards the optimum $\mathbf{x}_{\text{opt}}$, (ii) the vector of Lagrange factors γ_t towards the vector of Lagrange multipliers γ_{opt}, and (iii) the step-size σ_t towards 0.

7.1 Homogeneous Markov Chain

We start by recalling the definition of positive homogeneity.

Definition 5. [Definition 4 from [3]] A function $p : X \to Y$ is positive homogeneous of degree $k > 0$ with respect to $\mathbf{x}^* \in X$ if for all $\alpha > 0$ and for all $\mathbf{x} \in X$,

$$p(\mathbf{x}^* + \alpha \mathbf{x}) = \alpha^k p(\mathbf{x}^* + \mathbf{x}) \ . \tag{19}$$

Example 1. Our linear constraint functions, $g_i(\mathbf{x}) = \mathrm{b}_i^\intercal \mathbf{x} + c_i$, are positive homogeneous of degree 1 with respect to any $\mathbf{x}^* \in \mathbb{R}^n$ that satisfies $g_i(\mathbf{x}^*) = 0$. Indeed,

$$\begin{aligned} g_i(\mathbf{x}^* + \alpha \mathbf{x}) &= \mathrm{b}_i^\intercal(\mathbf{x}^* + \alpha \mathbf{x}) + c_i = \alpha(\mathrm{b}_i^\intercal \mathbf{x} + c_i) + \alpha \mathrm{b}_i^\intercal \mathbf{x}^* \\ &= \alpha g_i(\mathbf{x}^* + \mathbf{x}) \ , \text{ for all } \alpha > 0. \end{aligned} \tag{20}$$

The following theorem gives sufficient conditions under which the sequence $(\Phi_t)_{t \in \mathbb{N}}$, with $\Phi_t = (\mathbf{Y}_t, \Gamma_t, \omega_t)$, is a homogeneous Markov chain, where the random variables $\mathbf{Y}_t$ and Γ_t are defined in (21) below.

THEOREM 1. *Consider the $(\mu/\mu_{\text{w}}, \lambda)$-ES with augmented Lagrangian constraint handling minimizing the augmented Lagrangian h in (10), such that the step-size update function $\mathcal{G}_\sigma$ satisfies the condition in (17). Let $(\boldsymbol{X}_t, \sigma_t, \gamma_t, \omega_t)_{t \in \mathbb{N}}$ be the Markov chain associated to this ES and let $(\boldsymbol{U}_t)_{t \in \mathbb{N}}$ be a sequence of i.i.d. normal vectors. Let $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ such that $g_i(\bar{\boldsymbol{x}}) = 0$ for all $i = 1, \ldots, m$, and let $\bar{\gamma} \in \mathbb{R}^m$. Let*

$$\boldsymbol{Y}_t = \frac{\boldsymbol{X}_t - \bar{\boldsymbol{x}}}{\sigma_t} \ \text{ and } \ \Gamma_t = \frac{\gamma_t - \bar{\gamma}}{\sigma_t} \ . \tag{21}$$

Then, if the function $\mathcal{D}h_{\bar{\boldsymbol{x}}, \bar{\gamma}, \omega} : \mathbb{R}^{n+m} \to \mathbb{R}$ defined as

$$\mathcal{D}h_{\bar{\boldsymbol{x}}, \bar{\gamma}, \omega}(\boldsymbol{x}, \gamma) = h(\boldsymbol{x}, \gamma, \omega) - h(\bar{\boldsymbol{x}}, \bar{\gamma}, \omega) \tag{22}$$

is positive homogeneous of degree 2 with respect to $[\bar{\boldsymbol{x}}, \bar{\gamma}]$, then the sequence $(\Phi_t)_{t \in \mathbb{N}}$, where $\Phi_t = (\boldsymbol{Y}_t, \Gamma_t, \omega_t)$, is a homogeneous Markov chain that can be defined independently of $(\boldsymbol{X}_t, \sigma_t, \gamma_t, \omega_t)$ as $\boldsymbol{Y}_0 = (\boldsymbol{X}_0 - \bar{\boldsymbol{x}})/\sigma_0$, $\Gamma_0 = (\gamma_0 - \bar{\gamma})/\sigma_0$ and for all t

$$\boldsymbol{Y}_{t+1} = \mathcal{G}_{\boldsymbol{x}}((\boldsymbol{Y}_t, 1), \varsigma * \boldsymbol{U}_{t+1})/\mathcal{G}_\sigma(1, \varsigma * \boldsymbol{U}_{t+1}) \ , \tag{23}$$

$$\Gamma_{t+1}^i = \mathcal{H}_\gamma^{g_i(.+\bar{\boldsymbol{x}})}(\Gamma_t^i, \omega_t^i, \tilde{\boldsymbol{Y}}_{t+1})/\mathcal{G}_\sigma(1, \varsigma * \boldsymbol{U}_{t+1}) \ , \tag{24}$$

$$\omega_{t+1}^i = \mathcal{H}_\omega^{(f(.+\bar{\boldsymbol{x}}), g_i(.+\bar{\boldsymbol{x}}))}(\omega_t^i, \Gamma_t^i + \bar{\gamma}^i, \tilde{\boldsymbol{Y}}_{t+1}) \ , \tag{25}$$

with

$$\varsigma = \mathit{Ord}(h(\boldsymbol{Y}_t + \boldsymbol{U}_{t+1}^i + \bar{\boldsymbol{x}}, \Gamma_t + \bar{\gamma}, \omega_t)_{i=1,\cdots,\lambda}) \ , \tag{26}$$

$$\tilde{\boldsymbol{Y}}_{t+1} = \mathcal{G}_{\boldsymbol{x}}((\boldsymbol{Y}_t, 1), \varsigma * \boldsymbol{U}_{t+1}) \ , \tag{27}$$

where the Ord operator extracts the permutation of ordered candidate solutions (see (5)).

The proof of Theorem 1 is given in Appendix A.3. The key idea in the proof is that when $\mathcal{D}h_{\bar{\mathbf{x}}, \bar{\gamma}, \omega_t}$ is positive homogeneous of degree 2 with respect to $[\bar{\mathbf{x}}, \bar{\gamma}]$, the same permutation ς is obtained when ranking candidate solutions $\mathbf{X}_t + \sigma_t \mathbf{U}_{t+1}^i$ on $h(., \gamma_t, \omega_t)$ than when ranking candidate solutions $\mathbf{Y}_t + \mathbf{U}_{t+1}^i$ on $h(. + \bar{\mathbf{x}}, \Gamma_t + \bar{\gamma}, \omega_t)$, i.e.,

$$\varsigma_{(\mathbf{X}_t, \sigma_t)}^{h(., \gamma_t, \omega_t)} = \varsigma_{(\mathbf{Y}_t, 1)}^{h(. + \bar{\mathbf{x}}, \Gamma_t + \bar{\gamma}, \omega_t)} = \varsigma \ .$$

Scale-invariance of Algorithm 1, induced by the property of $\mathcal{G}_\sigma$ in (17), is also used explicitly in the proof while translation-invariance is used implicitly.

Theorem 1 holds for any $\bar{\mathbf{x}} \in \mathbb{R}^n$ such that $g_i(\bar{\mathbf{x}}) = 0$, for all $i \in \{1, \cdots, m\}$, and for any $\bar{\gamma} \in \mathbb{R}^m$. In particular, it holds for the optimum $\mathbf{x}_{\text{opt}}$ of our constrained problem and the associated vector γ_{opt} of Lagrange multipliers.

The following corollary states that on convex quadratic functions, $(\Phi_t)_{t \in \mathbb{N}}$ (defined in Theorem 1) is a homogeneous Markov chain for $\bar{\mathbf{x}} = \mathbf{x}_{\text{opt}}$ and $\bar{\gamma} = \gamma_{\text{opt}}$.

COROLLARY 1. *Let $(\boldsymbol{X}_t, \sigma_t, \gamma_t, \omega_t)_{t \in \mathbb{N}}$ be the Markov chain associated to the $(\mu/\mu_{\text{w}}, \lambda)$-ES in 1 optimizing the augmented Lagrangian h in (10), with f convex quadratic defined as*

$$f(\boldsymbol{x}) = \frac{1}{2} \boldsymbol{x}^T \boldsymbol{H} \boldsymbol{x} \ , \tag{28}$$

where $\boldsymbol{H} \in \mathbb{R}^{n \times n}$ is a symmetric positive-definite matrix. Let $\boldsymbol{Y}_t = \frac{\boldsymbol{X}_t - \boldsymbol{x}_{opt}}{\sigma_t}$ and $\Gamma_t = \frac{\gamma_t - \gamma_{\text{opt}}}{\sigma_t}$, where x_{opt} is the optimum of the constrained problem and γ_{opt} is the vector of the associated Lagrange multipliers. Then $(\Phi_t)_{t \in \mathbb{N}}$, with $\Phi_t = (\boldsymbol{Y}_t, \Gamma_t, \omega_t)$, is a homogeneous Markov chain defined independently of $(\boldsymbol{X}_t, \sigma_t, \gamma_t, \omega_t)$ as in (23), (24), (25), (26), and (27) by taking $\bar{\boldsymbol{x}} = \boldsymbol{x}_{opt}$ and $\bar{\gamma} = \gamma_{\text{opt}}$.

We prove the corollary by showing that the function $\mathcal{D}h_{\mathbf{x}_{\text{opt}}, \gamma_{\text{opt}}, \omega}$ defined in (22) is positive homogeneous of degree 2 with respect to $[\mathbf{x}_{\text{opt}}, \gamma_{\text{opt}}]$ for $f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^\intercal \mathbf{H} \mathbf{x}$. For the proof (see Appendix A.4), we use the following elements:

- The definitions of the gradients of f and g_i, $\nabla_{\mathbf{x}} f(\mathbf{y}) = \mathbf{y}^\intercal \mathbf{H}$ and $\nabla_{\mathbf{x}} g_i(\mathbf{y}) = \mathrm{b}_i^\intercal$, respectively.
- The KKT stationarity condition at the optimum $\mathbf{x}_{\text{opt}}$

$$\nabla_{\mathbf{x}} f(\mathbf{x}_{\text{opt}}) + \sum_{i=1}^m \gamma^i \nabla_{\mathbf{x}} g_i(\mathbf{x}_{\text{opt}}) = \mathbf{0} \ . \tag{29}$$

Remark 2. For a convex quadratic objective function f and linear constraints g_i, $i = 1, \cdots, m$, KKT conditions are sufficient conditions for optimality. That is, a point that satisfies KKT conditions is also an optimum of the constrained problem (see [15, Theorem 16.4]). The optimization problem we consider is unimodal, therefore $\mathbf{x}_{\text{opt}}$ is the only point satisfying the KKT conditions.

7.2 Sufficient Conditions for Linear Convergence

In the sequel, we investigate linear convergence of Algorithm 1. There exist many definitions—not always equivalent—of linear convergence. We consider here the almost sure linear convergence whose definition is given in Definition 1. We will also briefly discuss another definition of linear convergence that considers the expected log-progress $\ln \frac{\|\mathbf{X}_{t+1}-\mathbf{x}_{\text{opt}}\|}{\|\mathbf{X}_t-\mathbf{x}_{\text{opt}}\|}$.

We start by giving the definitions of an invariant probability measure and positivity [14]. We consider a Markov chain $(\mathbf{X}_t)_{t\in\mathbb{N}}$ that takes its values in a set $\mathcal{X} \subset \mathbb{R}^n$ equipped with its Borel σ-algebra $\mathcal{B}(\mathcal{X})$. The transition probabilities are given by the transition probability kernel P such that for $\mathbf{x} \in \mathcal{X}$ and $B \in \mathcal{B}(\mathcal{X})$

$$P(\mathbf{x}, B) = \Pr(\mathbf{X}_{t+1} \in B \mid \mathbf{X}_t = \mathbf{x}) \ .$$

Definition 6. Let π be a probability measure on $\mathcal{X}$ and let $\mathbf{X}_t \sim \pi$. We say that π is invariant if

$$\pi(B) = \int_{\mathcal{X}} \pi(d\mathbf{x}) P(\mathbf{x}, B) \ .$$

We say that a Markov chain is positive if there exists an invariant probability measure for this Markov chain.

Harris-recurrence [14] is related to the notion of irreducibility. Informally, a Markov chain is φ-irreducible if there exists a nonzero measure φ on $\mathcal{X}$ such that all φ-positive sets (that is, sets $B \in \mathcal{B}(\mathcal{X})$ such that $\varphi(B) > 0$) are reachable from anywhere in $\mathcal{X}$. In such a case, there exists a maximal irreducibility measure ψ that dominates other irreducibility measures [14].

Definition 7. Let $(\mathbf{X}_t)_{t\in\mathbb{N}}$ be a ψ-irreducible Markov chain. A measurable set $B \in \mathcal{B}(\mathcal{X})$ is Harris-recurrent if

$$\Pr\Big(\sum_{t\in\mathbb{N}_>} \mathbf{1}_{\{\mathbf{X}_t \in B\}} = \infty \mid \mathbf{X}_0 = \mathbf{x}\Big) = 1 \ ,$$

for all $\mathbf{x} \in B$. By extension, we say that $(\mathbf{X}_t)_{t\in\mathbb{N}}$ is Harris-recurrent if all ψ-positive sets are Harris-recurrent.

We can now recall Theorem 17.0.1 from [14] that gives sufficient conditions for the application of a LLN for Markov chains.

THEOREM 2 (THEOREM 17.0.1 FROM [14]). *Let $\mathbf{Z}$ be a positive Harris-recurrent chain with invariant probability π. Then the LLN holds for any function q such that $\pi(|q|) = \int |q(\boldsymbol{z})|\pi(d\boldsymbol{z}) < \infty$, that is, for any initial state $\mathbf{Z}_0$, $\lim_{t\to\infty} \frac{1}{t}\sum_{k=0}^{t-1} q(\mathbf{Z}_k) = \pi(q)$ almost surely.*

Consider now Algorithm 1 minimizing the augmented Lagrangian h in (10) corresponding to the optimization problem in (8), such that the function $\mathcal{D}h_{\mathbf{x}_{\text{opt}},\gamma_{\text{opt}},\omega_t}$ defined in (22) is positive homogeneous of degree 2 with respect to $[\mathbf{x}_{\text{opt}}, \gamma_{\text{opt}}]$. By virtue of Theorem 1, $(\Phi_t)_{t\in\mathbb{N}}$ is a homogeneous Markov chain. The following theorem gives sufficient conditions under which Algorithm 1 converges to the optimum $\mathbf{x}_{\text{opt}}$ of the constrained problem, as well as to the corresponding Lagrange multiplier γ_{opt}.

THEOREM 3. *Let $(\boldsymbol{X}_t, \sigma_t, \gamma_t, \omega_t)_{t\in\mathbb{N}}$ be the Markov chain associated to Algorithm 1 optimizing the augmented Lagrangian h such that the function $\mathcal{D}h_{\boldsymbol{x}_{opt},\gamma_{\text{opt}},\omega_t}$ defined in (22) is positive homogeneous of degree 2 with respect to $[\boldsymbol{x}_{opt}, \gamma_{\text{opt}}]$, where $\boldsymbol{x}_{opt}$ is the optimum of the constrained problem (8) and γ_{opt} is the corresponding Lagrange multiplier. Let $(\Phi_t)_{t\in\mathbb{N}}$ be the Markov chain defined in Theorem 1 and assume that it is positive Harris-recurrent with invariant probability measure π, that $E_\pi(|\ln\|[\phi]_1\|\|) < \infty$, $E_\pi(|\ln\|[\phi]_2\|\|) < \infty$, and $E_\pi(\mathcal{R}(\phi)) < \infty$, where*

$$\mathcal{R}(\phi) = E\big(\ln(\mathcal{G}_\sigma(1, \varsigma * \boldsymbol{U}_{t+1}))|\Phi_t = \phi\big) \ . \tag{30}$$

Then for all $\boldsymbol{X}_0$, for all σ_0, for all γ_0, and for all ω_0,

$$\lim_{t\to\infty} \frac{1}{t} \ln \frac{\|\boldsymbol{X}_t - \boldsymbol{x}_{opt}\|}{\|\boldsymbol{X}_0 - \boldsymbol{x}_{opt}\|} = \lim_{t\to\infty} \frac{1}{t} \ln \frac{\|\gamma_t - \gamma_{\text{opt}}\|}{\|\gamma_0 - \gamma_{\text{opt}}\|} = \lim_{t\to\infty} \frac{1}{t} \ln \frac{\sigma_t}{\sigma_0}$$
$$= -CR \ a.s. \ ,$$

where

$$-CR = \int \mathcal{R}(\phi)\pi(d\phi) \ .$$

The proof idea is similar to the one discussed in Section 3 for the unconstrained case, where the quantities $\frac{1}{t} \ln \frac{\|\mathbf{X}_t-\mathbf{x}_{\text{opt}}\|}{\|\mathbf{X}_0-\mathbf{x}_{\text{opt}}\|}$, $\frac{1}{t} \ln \frac{\|\gamma_t-\gamma_{\text{opt}}\|}{\|\gamma_0-\gamma_{\text{opt}}\|}$, and $\frac{1}{t} \ln \frac{\sigma_t}{\sigma_0}$ are expressed as a function of the Markov chain Φ_t. The detailed proof of Theorem 1 is given in Appendix A.5.

While in the previous theorem we have presented sufficient conditions on the Markov chain Φ_t for the almost sure linear convergence of the algorithm, other sufficient conditions can allow to derive the geometric convergence of the expected log-progress. Typically, assuming we have proven a so-called geometric drift for the chain Φ_t, plus some assumptions ensuring that the conditional log-progress is dominated by the drift function (see for instance [7, Theorem 5.4]), then

$$\sum_t r^t |E_{\Phi_0} \ln \frac{\|\mathbf{X}_{t+1} - \mathbf{x}_{\text{opt}}\|}{\|\mathbf{X}_t - \mathbf{x}_{\text{opt}}\|} - (-\text{CR})| \le RV(\Phi_0) \ , \tag{31}$$

where $r > 1$, R is a positive constant and $V \ge 1$ is the drift function. Equation (31) also holds when replacing $\ln \frac{\|\mathbf{X}_{t+1}-\mathbf{x}_{\text{opt}}\|}{\|\mathbf{X}_t-\mathbf{x}_{\text{opt}}\|}$ by $\ln \frac{\|\gamma_{t+1}-\gamma_{\text{opt}}\|}{\|\gamma_t-\gamma_{\text{opt}}\|}$ and $\ln \frac{\sigma_{t+1}}{\sigma_t}$.

8. EMPIRICAL RESULTS

We describe here our experimental setting and discuss the obtained results.

8.1 Step-Size Adaptation Mechanism

We test Algorithm 1 with cumulative step-size adaptation (CSA) [12]. The idea of CSA consists in keeping track of the successive steps taken by the algorithm in the search space. This is done by computing an evolution path, p_t, according to

$$p_{t+1} = (1 - c_\sigma)p_t + \sqrt{\frac{c_\sigma(2 - c_\sigma)}{\sum_{k=1}^{\mu} w_k^2}} \sum_{k=1}^{\mu} w_k \mathbf{U}_{t+1}^{\varsigma(k)} \ , \tag{32}$$

where $0 < c_\sigma \le 1$ and $\boldsymbol{p}_0^\sigma = \mathbf{0}$. The constant $\sqrt{\frac{c_\sigma(2-c_\sigma)}{\sum_{k=1}^{\mu} w_k^2}}$ is a normalization factor that is chosen such that under random selection, if p_t is normally distributed ($p_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{n\times n})$), then p_{t+1} is identically distributed [10, 11]. The evolution path is used to adapt the step-size σ_t according to the following rule.

$$\sigma_{t+1} = \sigma_t \exp^{\frac{c_\sigma}{d_\sigma}} \left(\frac{\|p_{t+1}\|}{E\|\mathcal{N}(\mathbf{0}, \mathbf{I}_{n\times n})\|} - 1 \right) \ . \tag{33}$$

The norm of the evolution path is compared to the expected norm of a standard normal vector by computing the ratio $\frac{\|p_{t+1}\|}{E\|\mathcal{N}(\mathbf{0},\mathbf{I}_{n\times n})\|}$ and the step-size is updated depending on this ratio: if $\frac{\|p_{t+1}\|}{E\|\mathcal{N}(\mathbf{0},\mathbf{I}_{n\times n})\|} \geq 1$, σ_t is increased as this suggests that the progress is too slow. Otherwise, σ_t is decreased. d_σ is a damping factor whose role is to moderate the changes in σ_t values.

In order for this adaptation mechanism to be compliant with our general adaptation rule $\mathcal{G}_\sigma(\sigma_t, \varsigma * \mathbf{U}_{t+1})$ (see (13)), we take $c_\sigma = 1$, that is, we consider CSA without cumulation. In this case, (32) becomes

$$p_{t+1} = \sqrt{\frac{1}{\sum_{k=1}^{\mu} w_k^2}} \sum_{k=1}^{\mu} w_k \mathbf{U}_{t+1}^{\varsigma(k)} \quad .$$

For the damping factor, we use

$$d_\sigma = 2 + 2\max\left(0, \sqrt{\frac{1/\sum_{k=1}^{\mu} w_k^2 - 1}{n+1}} - 1\right) \quad ,$$

which is the default value recommended in [11] with $c_\sigma = 1$.

8.2 Simulations of the Markov Chain and Single Runs

We test Algorithm 1 on two convex quadratic functions, as a particular case of Corollary 1: the sphere function, f_{sphere}, and the ellipsoid function, $f_{\text{ellipsoid}}$, with a moderate condition number. They are defined according to (28) by taking (i) $\mathbf{H} = \mathbf{I}_{n\times n}$ for f_{sphere} and (ii) $\mathbf{H}$ diagonal with diagonal elements $[\mathbf{H}]_i = \alpha^{\frac{i-1}{n-1}}$, $i = 1, \cdots, n$, for $f_{\text{ellipsoid}}$ and with a condition number $\alpha = 10$.

We choose $\mathbf{x}_{\text{opt}}$ to be at $(10, \cdots, 10)^\intercal$ and construct the (active) constraints following the steps below:

- For the first constraint, $\mathrm{b}_1 = -\nabla_{\mathbf{x}} f(\mathbf{x}_{\text{opt}})^\intercal$ and $c_1 = -\mathrm{b}_1^\intercal \mathbf{x}_{\text{opt}}$,
- For the $m-1$ remaining constraints, we choose the constraint normal b_i as a standard normal vector ($\mathrm{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{n\times n})$) and $c_i = -\mathrm{b}_i^\intercal \mathbf{x}_{\text{opt}}$. We choose the point $\nabla_{\mathbf{x}} f(\mathbf{x}_{\text{opt}})^\intercal = -\mathrm{b}_1$ to be feasible, along with $\mathbf{x}_{\text{opt}}$. Therefore, if $g_i(\nabla_{\mathbf{x}} f(\mathbf{x}_{\text{opt}})^\intercal) > 0$, we modify b_i and c_i according to: $\mathrm{b}_i = -\mathrm{b}_i$ and $c_i = -c_i$.

With the construction above, the constraints satisfy the LICQ (see Remark 1) with probability one. In such a case, the unique vector of Lagrange multipliers associated to $\mathbf{x}_{\text{opt}}$ is $\gamma_{\text{opt}} = (1, 0, \cdots, 0)^\intercal$.

As for the parameters of Algorithm 1, we choose the default values in [11] for both λ and μ. We set the weights w_i, $i = 1, \cdots, \mu$, according to [1], where they are chosen to be optimal on the sphere function in infinite dimension. We take $d_\gamma = d_\omega = 5$, $\chi = 2^{1/n}$, $k_1 = 3$, and $k_2 = 5$.

We run Algorithm 1 and simulate the Markov chain $(\Phi_t)_{t\in\mathbb{N}}$ (defined in Theorem 1) in $n = 10$ on f_{sphere} and $f_{\text{ellipsoid}}$ with $m = 1, 2, 5, 9$ constraints. For each problem, we test three different initial values of the penalty vector $\omega_0 = (1, \cdots, 1)^\intercal, (10^3, \cdots, 10^3)^\intercal$, $(10^{-3}, \cdots, 10^{-3})^\intercal$. In all the tests, $\mathbf{X}_0$ and $\mathbf{Y}_0$ are sampled uniformly in $[-5, 5]^n$, $\sigma_0 = 1$, and $\gamma_0 = \Gamma_0 = (5, \cdots, 5)^\intercal$.

Figures 2-5 show simulations of the Markov chain on f_{sphere} (left column) and $f_{\text{ellipsoid}}$ (right column) subject to 1, 2, 5, and 9 constraints respectively. Displayed are the normalized distance to $\mathbf{x}_{\text{opt}}$, $\|\mathbf{Y}_t\|$ (red), the normalized distance to γ_{opt}, $\|\Gamma_t\|$ (green), and the norm of the vector of penalty factors, $\|\omega_t\|$ (blue) in log-scale, for $\omega_0 = (1, \cdots, 1)^\intercal$ (first row), $\omega_0 = (10^3, \cdots, 10^3)^\intercal$ (second row), and $\omega_0 = (10^{-3}, \cdots, 10^{-3})^\intercal$ (third row). We observe an overall convergence to a stationary distribution, independently of ω_0, after a certain number of iterations. For $\omega_0 = (10^3, \cdots, 10^3)^\intercal$,

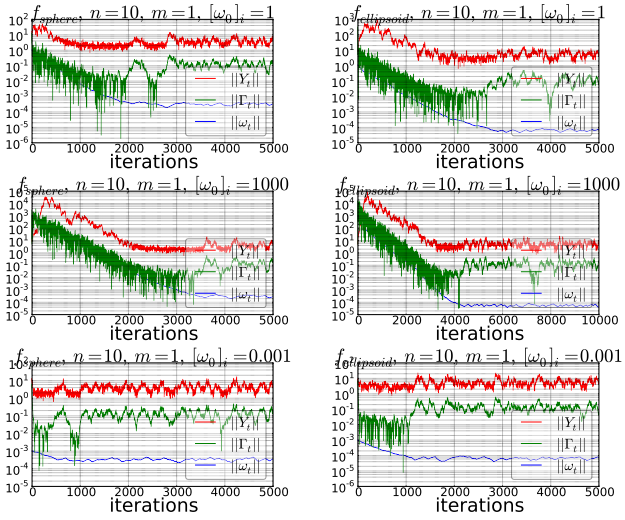

Figure 2: Simulations of the Markov chain on f_{sphere} (left) and $f_{\text{ellipsoid}}$ (right) with $m = 1$ in $n = 10$.

the adaptation phase before reaching the stationary state is longer than with $\omega_0 = (1, \cdots, 1)^\intercal$ or $\omega_0 = (10^{-3}, \cdots, 10^{-3})^\intercal$ on both f_{sphere} and $f_{\text{ellipsoid}}$. It also increases with increasing m: it takes approximately 4×10^3 iterations on f_{sphere} and $f_{\text{ellipsoid}}$ with $m = 1$ (Figure 2) and approximately 6×10^3 iterations with $m = 9$ (Figure 5). Indeed, the problem becomes more difficult for large m (as shown below with single runs). We also observe from Figures 2-5 that $\|\omega_t\|$ stabilizes around a larger value as m increases (approximately 4×10^{-4} and 6×10^{-5} on f_{sphere} and $f_{\text{ellipsoid}}$ respectively with $m = 1$ versus approximately 1 and 4 with $m = 9$).

Figures 6-9 show single runs of Algorithm 1 on the same constrained problems described previously. Results on constrained f_{sphere} and constrained $f_{\text{ellipsoid}}$ are displayed in left and right columns respectively. The displayed quantities are (i) the distance to the optimum, $\|\mathbf{X}_t - \mathbf{x}_{\text{opt}}\|$ (red), (ii) the distance to the Lagrange multipliers, $\|\gamma_t - \gamma_{\text{opt}}\|$ (green), (iii) the norm of the penalty vector, $\|\omega_t\|$ (blue), and (iv) the step-size, σ_t (purple), in log-scale. Linear convergence occurs after an adaptation phase whose length depends on the accuracy of the choice of the initial parameters: for $m = 1$ and $\omega_0 = (10^{-3}, \cdots, 10^{-3})^\intercal$ (Figure 6, third row), linear convergence occurs after only around 30 iterations because ω_0 is already close to a stationary value (see Figure 2). On f_{sphere} with $m = 2$ (Figure 7, left column), the algorithm reaches a distance to $\mathbf{x}_{\text{opt}}$ of 10^{-4} after around 750 iterations with $\omega_0 = (1, \cdots, 1)^\intercal$, compared to around 2500 iterations with $\omega_0 = (10^3, \cdots, 10^3)^\intercal$ and around 1300 iterations with $\omega_0 = (10^{-3}, \cdots, 10^{-3})^\intercal$. The reason is that $\omega_0 = (1, \cdots, 1)^\intercal$ is closer to the stationary value in this case (Figure 3, left column). As the number of constraints increases (Figures 8 and 9), the number of iterations needed to reach a given precision increases: it takes more than 2 times longer to reach a distance from the optimum of 10^{-4} on both f_{sphere} and $f_{\text{ellipsoid}}$ with $m = 9$ and $\omega_0 = (1, \cdots, 1)^\intercal$ (Figure 9, first row) than with $m = 1$ (Figure 6, first row). These results are consistent with the simulations of the Markov chain in that the observed stability of the Markov chain leads to linear convergence (or divergence) of the algorithm—as stated in Theorem 3.

9. DISCUSSION

In this work, we investigated linear convergence of a $(\mu/\mu_{\text{w}}, \lambda)$-ES with an augmented Lagrangian constraint handling on the lin-

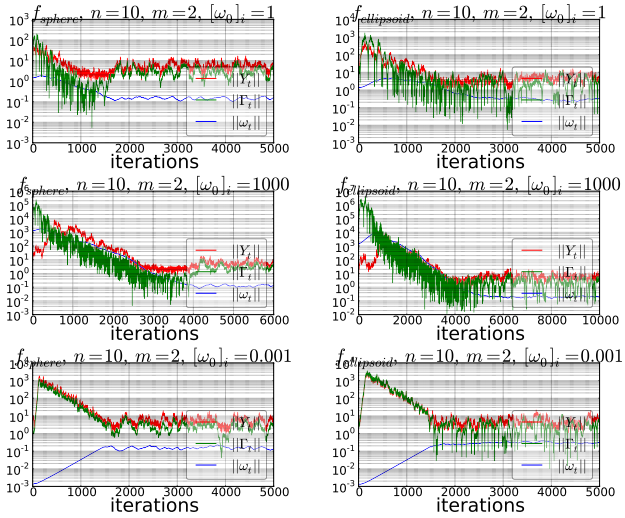

Figure 3: Simulations of the Markov chain on f_{sphere} (left) and $f_{\text{ellipsoid}}$ (right) with $m = 2$ in $n = 10$.

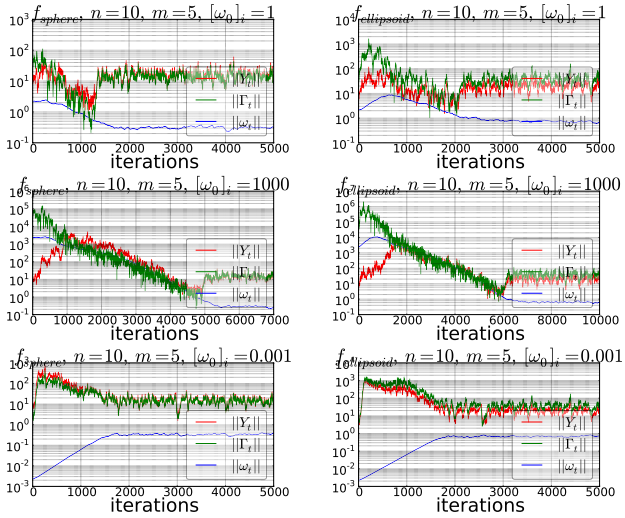

Figure 4: Simulations of the Markov chain on f_{sphere} (left) and $f_{\text{ellipsoid}}$ (right) with $m = 5$ in $n = 10$.

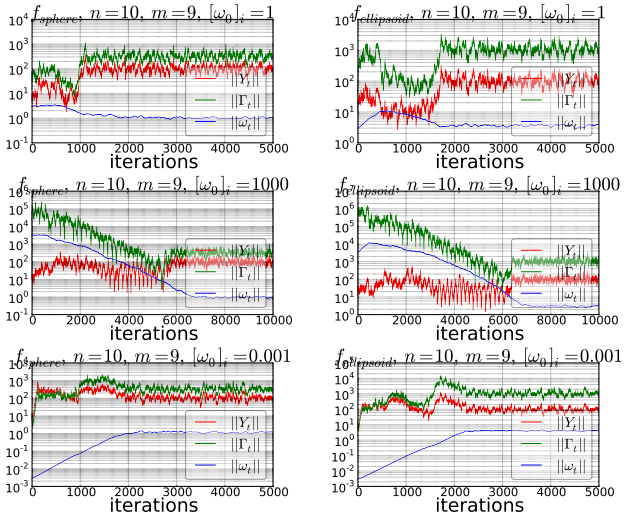

Figure 5: Simulations of the Markov chain on f_{sphere} (left) and $f_{\text{ellipsoid}}$ (right) with $m = 9$ in $n = 10$.

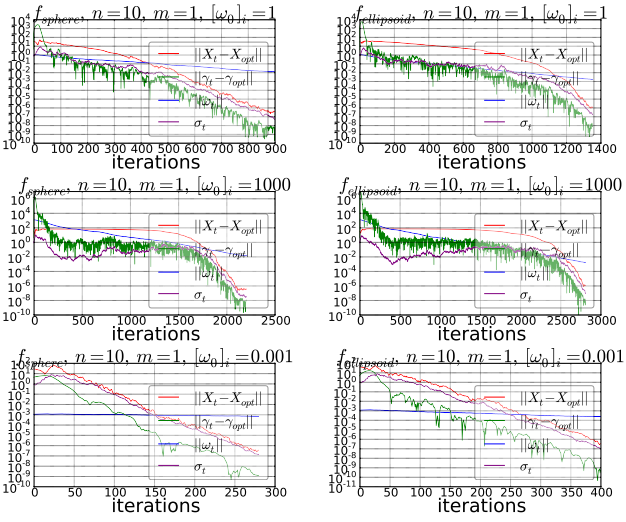

Figure 6: Single runs on f_{sphere} (left) and $f_{\text{ellipsoid}}$ (right) with $m = 1$ in $n = 10$, with three different values of ω_0.

early constrained problem where all the constraints are active. We adopted a Markov chain approach and exhibited a homogeneous Markov chain on problems where the associated augmented Lagrangian, centered in the optimum and the corresponding Lagrange multipliers, is positive homogeneous of degree 2. We gave sufficient stability conditions which, when satisfied by the Markov chain, lead to linear convergence to the optimum as well as to the Lagrange multipliers. Simulations of the Markov chain on linearly constrained convex quadratic functions (as a particular case of the exhibited class of functions) show empirical evidence of stability for the tested parameter setting. We draw attention, however, to the fact that the observed stability may depend on the chosen parameter setting—in particular the damping factors for the Lagrange factors and the penalty factors—and proper parameter values are necessary to observe stability, especially in larger dimensions and for large numbers of constraints.

The conducted analysis gives insight into the behavior of the practical $(\mu/\mu_{\text{w}}, \lambda)$-ES obtained when optimizing the augmented Lagrangian presented in (9). Indeed, we focus our study on the most difficult case in practice, where all the constraints are active at the optimum.

Finally, this work illustrates how the Markov chain approach—which is already applied to prove linear convergence of randomized optimization algorithms in the unconstrained case—can be extended to the constrained case.

Acknowledgments

This work was supported by the grant ANR-2012-MONU-0009 (NumBBO) from the French National Research Agency.

10. REFERENCES

[1] D. V. Arnold. Optimal weighted recombination. In *Foundations of Genetic Algorithms*, pages 215–237. Springer, 2005.

[2] D. V. Arnold and J. Porter. Towards an Augmented Lagrangian Constraint Handling Approach for the

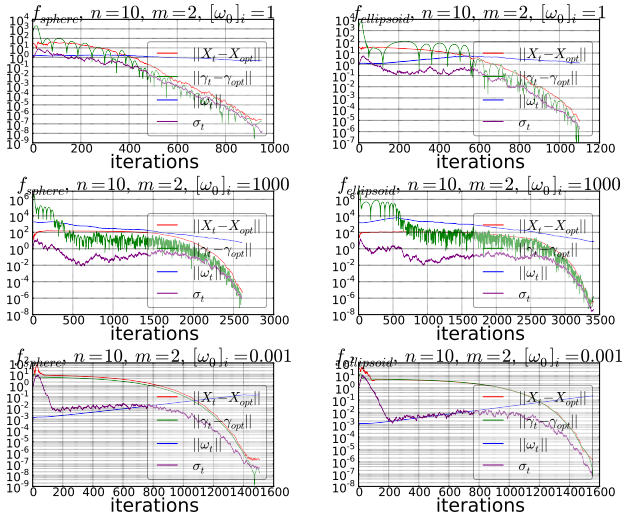

Figure 7: Single runs on f_{sphere} (left) and $f_{\text{ellipsoid}}$ (right) with $m = 2$ in $n = 10$, with three different values of ω_0.

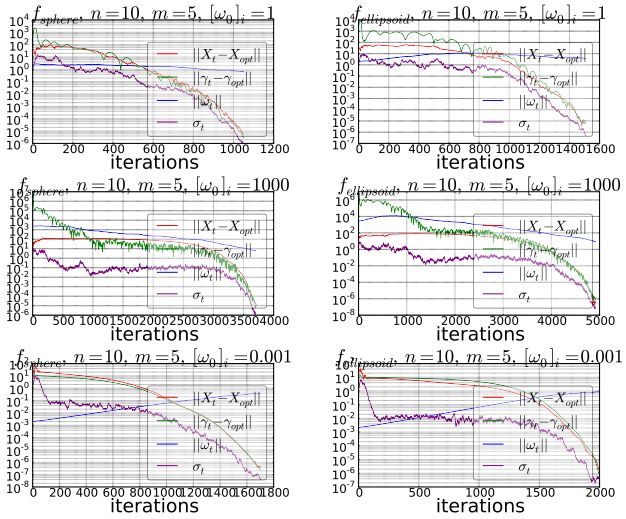

Figure 8: Single runs on f_{sphere} (left) and $f_{\text{ellipsoid}}$ (right) with $m = 5$ in $n = 10$, with three different values of ω_0.

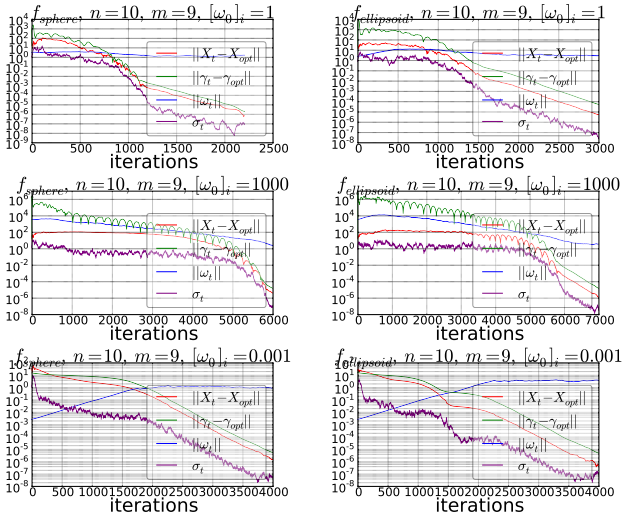

Figure 9: Single runs on f_{sphere} (left) and $f_{\text{ellipsoid}}$ (right) with $m = 9$ in $n = 10$, with three different values of ω_0.

$(1 + 1)$-ES. In *Genetic and Evolutionary Computation Conference*, pages 249–256. ACM Press, 2015.

[3] A. Atamna, A. Auger, and N. Hansen. Analysis of Linear Convergence of a $(1 + 1)$-ES with Augmented Lagrangian Constraint Handling. In *Genetic and Evolutionary Computation Conference*, pages 213–220. ACM Press, 2016.

[4] A. Atamna, A. Auger, and N. Hansen. Augmented Lagrangian Constraint Handling for CMA-ES—Case of a Single Linear Constraint. In *Parallel Problem Solving from Nature*, pages 181–191. Springer, 2016.

[5] A. Auger. Convergence Results for the $(1, \lambda)$-SA-ES Using the Theory of φ-Irreducible Markov Chains. *Theoretical Computer Science*, 334(1-3):35–69, 2005.

[6] A. Auger and N. Hansen. Linear Convergence on Positively Homogeneous Functions of a Comparison Based Step-Size Adaptive Randomized Search: the $(1 + 1)$ ES with Generalized One-Fifth Success Rule. Submitted for publication, 2013.

[7] A. Auger and N. Hansen. Linear Convergence of Comparison-Based Step-Size Adaptive Randomized Search via Stability of Markov Chains. *SIAM Journal on Optimization*, 26(3):1589–1624, 2016.

[8] A. Bienvenüe and O. François. Global Convergence of Evolution Strategies in Spherical Problems: Some Simple Proofs and Difficulties. *Theoretical Computer Science*, 306(1–3):269–289, 2003.

[9] K. Deb and S. Srivastava. A Genetic Algorithm Based Augmented Lagrangian Method for Constrained Optimization. *Computational Optimization and Applications*, 53(3):869–902, 2012.

[10] H. Hansen, D. V. Arnold, and A. Auger. Evolution strategies. In J. Kacprzyk and W. Pedrycz, editors, *Handbook of Computational Intelligence*, chapter 44, pages 871–898. Springer, 2015.

[11] N. Hansen. The CMA Evolution Strategy: A Tutorial. http://arxiv.org/pdf/1604.00772v1.pdf, 2016.

[12] N. Hansen and A. Ostermeier. Completely Derandomized Self-Adaptation in Evolution Strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

[13] M. R. Hestenes. Multiplier and Gradient Methods. *Journal of Optimization Theory and Applications*, 4(5):303–320, 1969.

[14] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, 1993.

[15] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 2nd edition, 2006.

[16] M. J. D. Powell. A Method for Nonlinear Constraints in Minimization Problems. In R. Fletcher, editor, *Optimization*, pages 283–298. Academic Press, 1969.

[17] M.-J. Tahk and B.-C. Sun. Coevolutionary Augmented Lagrangian Methods for Constrained Optimization. *IEE Transactions on Evolutionary Computation*, 4(2):114–124, 2000.

APPENDIX

A. PROOFS

A.1 Proof of Proposition 1

For Algorithm 1, the state $\mathbf{s}_t = (\mathbf{X}_t, \sigma_t, \gamma_t, \omega_t)$. Let

$$(\mathbf{X}'_{t+1}, \sigma'_{t+1}, \gamma'_{t+1}, \omega'_{t+1}) = \mathcal{F}^{(f(.-\mathbf{x}_0),\{g_i(.-\mathbf{x}_0)\}_{i=1,\cdots,m})}(\Phi(\mathbf{x}_0)(\mathbf{X}_t, \sigma_t, \gamma_t, \omega_t), \mathbf{U}_{t+1}) \ .$$

Given the definition of $\Phi(\mathbf{x}_0)$ in (16) and the update functions $\mathcal{G}_\mathbf{x}$, $\mathcal{G}_\sigma$, $\mathcal{H}_\gamma$, and $\mathcal{H}_\omega$ in (12), (13), (14), and (15) respectively, we have

$$\begin{aligned}\mathbf{X}'_{t+1} &= \mathcal{G}_\mathbf{x}((\mathbf{X}_t + \mathbf{x}_0, \sigma_t), \varsigma^{h(.-\mathbf{x}_0,\gamma_t,\omega_t)}_{(\mathbf{X}_t+\mathbf{x}_0,\sigma_t)} * \mathbf{U}_{t+1}) \\ &= \mathbf{X}_t + \mathbf{x}_0 + \sigma_t \sum_{i=1}^{\mu} w_i [\varsigma^{h(.-\mathbf{x}_0,\gamma_t,\omega_t)}_{(\mathbf{X}_t+\mathbf{x}_0,\sigma_t)} * \mathbf{U}_{t+1}]_i \ , \\ \sigma'_{t+1} &= \mathcal{G}_\sigma(\sigma_t, \varsigma^{h(.-\mathbf{x}_0,\gamma_t,\omega_t)}_{(\mathbf{X}_t+\mathbf{x}_0,\sigma_t)}) \ .\end{aligned}$$

On the other hand, we have

$$\begin{aligned}\varsigma^{h(.-\mathbf{x}_0,\gamma_t,\omega_t)}_{(\mathbf{X}_t+\mathbf{x}_0,\sigma_t)} &= Ord(h(\mathbf{X}_t + \mathbf{x}_0 + \sigma_t \mathbf{U}^i_{t+1} - \mathbf{x}_0, \gamma_t, \omega_t)_{i=1,\cdots,\lambda}) \\ &= \varsigma^{h(.,\gamma_t,\omega_t)}_{(\mathbf{X}_t,\sigma_t)} \ .\end{aligned}$$

It follows that

$$\begin{aligned}\mathbf{X}'_{t+1} &= \mathcal{G}_\mathbf{x}((\mathbf{X}_t, \sigma_t), \varsigma^{h(.,\gamma_t,\omega_t)}_{(\mathbf{X}_t,\sigma_t)} * \mathbf{U}_{t+1}) + \mathbf{x}_0 \\ &= \mathbf{X}_{t+1} + \mathbf{x}_0 \ , \\ \sigma'_{t+1} &= \mathcal{G}_\sigma(\sigma_t, \varsigma^{h(.,\gamma_t,\omega_t)}_{(\mathbf{X}_t,\sigma_t)}) = \sigma_{t+1} \ .\end{aligned} \tag{34}$$

Using (34), we obtain

$$\begin{aligned}\gamma'^i_{t+1} &= \mathcal{H}^{g_i(.-\mathbf{x}_0)}_\gamma(\gamma^i_t, \omega^i_t, \mathbf{X}'_{t+1}) = \gamma^i_t + \frac{\omega^i_t}{d_\gamma} g_i(\mathbf{X}'_{t+1} - \mathbf{x}_0) \\ &= \mathcal{H}^{g_i}_\gamma(\gamma^i_t, \omega^i_t, \mathbf{X}_{t+1}) = \gamma^i_{t+1}, \ i = 1, \cdots, m \ ,\end{aligned}$$

$$\begin{aligned}\omega'^i_{t+1} &= \mathcal{H}^{(f(.-\mathbf{x}_0),g_i(.-\mathbf{x}_0))}_\omega(\omega^i_t, \gamma^i_t, \mathbf{X}_t + \mathbf{x}_0, \mathbf{X}'_{t+1}) \\ &= \begin{cases} \omega^i_t \chi^{1/(4d_\omega)} & \text{if } \omega^i_t g_i(\mathbf{X}'_{t+1} - \mathbf{x}_0)^2 < \\ & \quad k_1 \frac{|h(\mathbf{X}'_{t+1}-\mathbf{x}_0,\gamma_t,\omega_t) - h(\mathbf{X}_t+\mathbf{x}_0-\mathbf{x}_0,\gamma_t,\omega_t)|}{n} \\ & \text{or } k_2 |g_i(\mathbf{X}'_{t+1} - \mathbf{x}_0) - g_i(\mathbf{X}_t + \mathbf{x}_0 - \mathbf{x}_0)| < \\ & \quad |g_i(\mathbf{X}_t + \mathbf{x}_0 - \mathbf{x}_0)| \\ \omega^i_t \chi^{-1/d_\omega} & \text{otherwise, } \ i = 1, \cdots, m \end{cases} \\ &= \mathcal{H}^{(f,g_i)}_\omega(\omega^i_t, \gamma^i_t, \mathbf{X}_t, \mathbf{X}_{t+1}) = \omega^i_{t+1}, \ i = 1, \cdots, m \ .\end{aligned}$$

Therefore,

$$(\mathbf{X}_{t+1} + \mathbf{x}_0, \sigma_{t+1}, \gamma_{t+1}, \omega_{t+1}) = \mathcal{F}^{(f(.-\mathbf{x}_0),\{g_i(.-\mathbf{x}_0)\}_{i=1,\cdots,m})}(\Phi(\mathbf{x}_0)(\mathbf{X}_t, \sigma_t, \gamma_t, \omega_t), \mathbf{U}_{t+1}) \ . \tag{35}$$

By applying the inverse transformation $\Phi(-\mathbf{x}_0)$ to (35), we recover $\mathcal{F}^{(f,\{g_i\}_{i=1,\cdots,m})}(\mathbf{X}_t, \sigma_t, \gamma_t, \omega_t)$.

A.2 Proof of Proposition 2

The state at iteration t is $\mathbf{s}_t = (\mathbf{X}_t, \sigma_t, \gamma_t, \omega_t)$. Let

$$(\mathbf{X}'_{t+1}, \sigma'_{t+1}, \gamma'_{t+1}, \omega'_{t+1}) = \mathcal{F}^{(f(\alpha.),\{g_i(\alpha.)\}_{i=1,\cdots,m})}(\Phi(\alpha)(\mathbf{X}_t, \sigma_t, \gamma_t, \omega_t), \mathbf{U}_{t+1}) \ .$$

By definition, we have

$$\begin{aligned}\varsigma^{h(\alpha.,\gamma_t,\omega_t)}_{(\mathbf{X}_t/\alpha,\sigma_t/\alpha)} &= Ord(h(\alpha(\mathbf{X}_t/\alpha + \sigma_t/\alpha \mathbf{U}^i_{t+1}), \gamma_t, \omega_t)_{i=1,\cdots,\lambda}) \\ &= \varsigma^{h(.,\gamma_t,\omega_t)}_{(\mathbf{X}_t,\sigma_t)} \ .\end{aligned}$$

Using the definition of $\Phi(\alpha)$ in (18), (12), (13), (14), (15), and the equation above, it follows

$$\begin{aligned}\mathbf{X}'_{t+1} &= \mathcal{G}_\mathbf{x}((\mathbf{X}_t/\alpha, \sigma_t/\alpha), \varsigma^{h(.,\gamma_t,\omega_t)}_{(\mathbf{X}_t,\sigma_t)} * \mathbf{U}_{t+1}) \\ &= \frac{\mathbf{X}_t}{\alpha} + \frac{\sigma_t}{\alpha} \sum_{i=1}^{\mu} w_i [\varsigma^{h(.,\gamma_t,\omega_t)}_{(\mathbf{X}_t,\sigma_t)} * \mathbf{U}_{t+1}]_i \\ &= \frac{1}{\alpha} \mathcal{G}_\mathbf{x}((\mathbf{X}_t, \sigma_t), \varsigma^{h(.,\gamma_t,\omega_t)}_{(\mathbf{X}_t,\sigma_t)} * \mathbf{U}_{t+1}) = \frac{\mathbf{X}_{t+1}}{\alpha} \ ,\end{aligned} \tag{36}$$

and $\sigma'_{t+1} = \mathcal{G}_\sigma(\sigma_t/\alpha, \varsigma^{h(.,\gamma_t,\omega_t)}_{(\mathbf{X}_t,\sigma_t)} * \mathbf{U}_{t+1})$. Using the scale-invariance property of $\mathcal{G}_\sigma$ (see (17)), we obtain

$$\sigma'_{t+1} = \frac{1}{\alpha} \mathcal{G}_\sigma(\sigma_t, \varsigma^{h(.,\gamma_t,\omega_t)}_{(\mathbf{X}_t,\sigma_t)} * \mathbf{U}_{t+1}) = \frac{\sigma_{t+1}}{\alpha} \ .$$

Finally, using (36) we get

$$\begin{aligned}\gamma'^i_{t+1} &= \mathcal{H}^{g_i(\alpha.)}_\gamma(\gamma^i_t, \omega^i_t, \mathbf{X}'_{t+1}) = \gamma^i_t + \frac{\omega^i_t}{d_\gamma} g_i(\alpha \mathbf{X}'_{t+1}) \\ &= \mathcal{H}^{g_i}_\gamma(\gamma^i_t, \omega^i_t, \mathbf{X}_{t+1}) = \gamma^i_{t+1}, \ i = 1, \cdots, m \ ,\end{aligned}$$

and

$$\begin{aligned}\omega'^i_{t+1} &= \mathcal{H}^{(f(\alpha.),g_i(\alpha.))}_\omega(\omega^i_t, \gamma^i_t, \mathbf{X}_t/\alpha, \mathbf{X}'_{t+1}) \\ &= \begin{cases} \omega^i_t \chi^{1/(4d_\omega)} & \text{if } \omega^i_t g_i(\alpha \mathbf{X}'_{t+1})^2 < \\ & \quad k_1 \frac{|h(\alpha \mathbf{X}'_{t+1},\gamma_t,\omega_t) - h(\alpha \mathbf{X}_t/\alpha,\gamma_t,\omega_t)|}{n} \\ & \text{or } k_2 |g_i(\alpha \mathbf{X}'_{t+1}) - g_i(\alpha \mathbf{X}_t/\alpha)| < \\ & \quad |g_i(\alpha \mathbf{X}_t/\alpha)| \\ \omega^i_t \chi^{-1/d_\omega} & \text{otherwise, } \ i = 1, \cdots, m \end{cases} \\ &= \mathcal{H}^{(f,g_i)}_\omega(\omega^i_t, \gamma^i_t, \mathbf{X}_t, \mathbf{X}_{t+1}) = \omega^i_{t+1}, \ i = 1, \cdots, m \ .\end{aligned}$$

Therefore,

$$\left(\frac{\mathbf{X}_{t+1}}{\alpha}, \frac{\sigma_{t+1}}{\alpha}, \gamma_{t+1}, \omega_{t+1}\right) = \mathcal{F}^{(f(\alpha.),\{g_i(\alpha.)\}_{i=1,\cdots,m})}(\Phi(\alpha)(\mathbf{X}_t, \sigma_t, \gamma_t, \omega_t), \mathbf{U}_{t+1}) \ . \tag{37}$$

By applying the inverse transformation $\Phi(1/\alpha)$ to (37), we obtain $\mathcal{F}^{(f,\{g_i\}_{i=1,\cdots,m})}(\mathbf{X}_t, \sigma_t, \gamma_t, \omega_t)$.

A.3 Proof of Theorem 1

We have

$$\mathbf{Y}_{t+1} = \frac{\mathbf{X}_{t+1} - \bar{\mathbf{x}}}{\sigma_{t+1}} = \frac{\mathcal{G}_\mathbf{x}((\mathbf{X}_t, \sigma_t), \varsigma * \mathbf{U}_{t+1}) - \bar{\mathbf{x}}}{\mathcal{G}_\sigma(\sigma_t, \varsigma * \mathbf{U}_{t+1})} \ .$$

Using translation-invariance and scale-invariance of Algorithm 1, it follows

$$\mathbf{Y}_{t+1} = \frac{\mathcal{G}_\mathbf{x}((\mathbf{Y}_t, 1), \varsigma * \mathbf{U}_{t+1})}{\mathcal{G}_\sigma(1, \varsigma * \mathbf{U}_{t+1})} \ ,$$

with

$$\begin{aligned}\varsigma &= \varsigma^{h(.,\gamma_t,\omega_t)}_{(\mathbf{X}_t,\sigma_t)} = \varsigma^{h(\sigma_t.+\bar{\mathbf{x}},\gamma_t,\omega_t)}_{(\mathbf{Y}_t,1)} = \varsigma^{h(\sigma_t.+\bar{\mathbf{x}},\sigma_t\Gamma_t+\bar{\gamma},\omega_t)}_{(\mathbf{Y}_t,1)} \\ &= Ord(h(\sigma_t(\mathbf{Y}_t + [\mathbf{U}_{t+1}]_i) + \bar{\mathbf{x}}, \sigma_t\Gamma_t + \bar{\gamma}, \omega_t)_{i=1,\cdots,\lambda}) \\ &= Ord(\mathcal{D}h_{\bar{\mathbf{x}},\bar{\gamma},\omega_t}(\sigma_t(\mathbf{Y}_t + [\mathbf{U}_{t+1}]_i) + \bar{\mathbf{x}}, \sigma_t\Gamma_t + \bar{\gamma})_{i=1,\cdots,\lambda}) \ ,\end{aligned}$$

where $\mathcal{D}h_{\bar{\mathbf{x}},\bar{\gamma},\omega_t}$ is defined in (22) and $\mathbf{Y}_t$ and Γ_t in (21). By positive homogeneity of $\mathcal{D}h_{\bar{\mathbf{x}},\bar{\gamma},\omega_t}$, it follows

$$\begin{aligned}\varsigma &= Ord(\sigma_t{}^2\mathcal{D}h_{\bar{\mathbf{x}},\bar{\gamma},\omega_t}(\mathbf{Y}_t + [\mathbf{U}_{t+1}]_i + \bar{\mathbf{x}}, \Gamma_t + \bar{\gamma})_{i=1,\cdots,\lambda}) \\ &= Ord(\mathcal{D}h_{\bar{\mathbf{x}},\bar{\gamma},\omega_t}(\mathbf{Y}_t + [\mathbf{U}_{t+1}]_i + \bar{\mathbf{x}}, \Gamma_t + \bar{\gamma})_{i=1,\cdots,\lambda}) \\ &= Ord(h(\mathbf{Y}_t + [\mathbf{U}_{t+1}]_i + \bar{\mathbf{x}}, \Gamma_t + \bar{\gamma}, \omega_t)_{i=1,\cdots,\lambda}) \\ &= \varsigma_{(\mathbf{Y}_t,1)}^{h(.+\bar{\mathbf{x}},\Gamma_t+\bar{\gamma},\omega_t)} \quad .\end{aligned}$$

On the other hand, we have

$$\begin{aligned}\Gamma_{t+1}^i &= \frac{\gamma_{t+1}^i - \bar{\gamma}^i}{\sigma_{t+1}} = \frac{\mathcal{H}_\gamma^{g_i}(\gamma_t^i, \omega_t^i, \mathbf{X}_{t+1}) - \bar{\gamma}^i}{\mathcal{G}_\sigma(\sigma_t, \varsigma * \mathbf{U}_{t+1})} \\ &= \frac{\gamma_t^i + \frac{\omega_t^i}{d_\gamma} g_i(\mathbf{X}_{t+1}) - \bar{\gamma}^i}{\sigma_t \mathcal{G}_\sigma(1, \varsigma * \mathbf{U}_{t+1})} = \frac{\Gamma_t^i + \frac{\omega_t^i}{d_\gamma \sigma_t} g_i(\mathbf{X}_{t+1})}{\mathcal{G}_\sigma(1, \varsigma * \mathbf{U}_{t+1})} \quad .\end{aligned}$$

Using positive homogeneity of g_i with respect to $\bar{\mathbf{x}}$ (see (20)) and the definition of $\tilde{\mathbf{Y}}_{t+1}$ in (27), we have

$$\begin{aligned}g_i(\mathbf{X}_{t+1}) &= g_i(\sigma_{t+1}\mathbf{Y}_{t+1} + \bar{\mathbf{x}}) \\ &= \sigma_t g_i(\underbrace{\mathcal{G}_\sigma(1, \varsigma * \mathbf{U}_{t+1})\mathbf{Y}_{t+1}}_{\mathcal{G}_{\mathbf{x}}((\mathbf{Y}_t,1),\varsigma*\mathbf{U}_{t+1})=\tilde{\mathbf{Y}}_{t+1}} + \bar{\mathbf{x}}) \quad . \qquad (38)\end{aligned}$$

Therefore,

$$\Gamma_{t+1}^i = \frac{\Gamma_t^i + \frac{\omega_t^i}{d_\gamma} g_i(\tilde{\mathbf{Y}}_{t+1} + \bar{\mathbf{x}})}{\mathcal{G}_\sigma(1, \varsigma * \mathbf{U}_{t+1})} = \frac{\mathcal{H}_\gamma^{g_i(.+\bar{\mathbf{x}})}(\Gamma_t^i, \omega_t^i, \tilde{\mathbf{Y}}_{t+1})}{\mathcal{G}_\sigma(1, \varsigma * \mathbf{U}_{t+1})} \quad ,$$

for $i = 1, \cdots, m$. Finally,

$$\begin{aligned}\omega_{t+1}^i &= \mathcal{H}_\omega^{(f,g_i)}(\omega_t^i, \gamma_t^i, \mathbf{X}_t, \mathbf{X}_{t+1}) \\ &= \begin{cases} \omega_t^i \chi^{1/(4d_\omega)} & \text{if } \omega_t^i g_i(\mathbf{X}_{t+1})^2 < \\ & k_1 \frac{|h(\mathbf{X}_{t+1},\gamma_t,\omega_t) - h(\mathbf{X}_t,\gamma_t,\omega_t)|}{n} \\ & \text{or } k_2 |g_i(\mathbf{X}_{t+1}) - g_i(\mathbf{X}_t)| < \\ & \qquad |g_i(\mathbf{X}_t)| \\ \omega_t^i \chi^{-1/d_\omega} & \text{otherwise, } i = 1, \cdots, m \end{cases} \\ &= \begin{cases} \omega_t^i \chi^{1/(4d_\omega)} & \text{if } \omega_t^i g_i(\tilde{\mathbf{Y}}_{t+1} + \bar{\mathbf{x}})^2 < \\ & k_1 \frac{|h(\tilde{\mathbf{Y}}_{t+1}+\bar{\mathbf{x}},\Gamma_t+\bar{\gamma},\omega_t) - h(\mathbf{Y}_t+\bar{\mathbf{x}},\Gamma_t+\bar{\gamma},\omega_t)|}{n} \\ & \text{or } k_2 |g_i(\tilde{\mathbf{Y}}_{t+1} + \bar{\mathbf{x}}) - g_i(\mathbf{Y}_t + \bar{\mathbf{x}})| < \\ & \qquad |g_i(\mathbf{Y}_t + \bar{\mathbf{x}})| \\ \omega_t^i \chi^{-1/d_\omega} & \text{otherwise} \end{cases} \\ &= \mathcal{H}_\omega^{(f(.+\bar{\mathbf{x}}),g_i(.+\bar{\mathbf{x}}))}(\omega_t^i, \Gamma_t^i + \bar{\gamma}^i, \mathbf{Y}_t, \tilde{\mathbf{Y}}_{t+1}) \quad ,\end{aligned}$$

for $i = 1, \cdots, m$, where we used (20), along with (38), and positive homogeneity of $\mathcal{D}h_{\bar{\mathbf{x}},\bar{\gamma},\omega_t}$ with respect to $[\bar{\mathbf{x}}, \bar{\gamma}]$ to deduce that

$$\begin{aligned}&h(\mathbf{X}_{t+1}, \gamma_t, \omega_t) - h(\mathbf{X}_t, \gamma_t, \omega_t) = \sigma_t^2(\mathcal{D}h_{\bar{\mathbf{x}},\bar{\gamma},\omega_t}(\tilde{\mathbf{Y}}_{t+1} + \bar{\mathbf{x}}, \sigma_t\Gamma_t + \bar{\gamma}) \\ &- \mathcal{D}h_{\bar{\mathbf{x}},\bar{\gamma},\omega_t}(\mathbf{Y}_{t+1} + \bar{\mathbf{x}}, \sigma_t\Gamma_t + \bar{\gamma})) \\ &= \sigma_t^2(h(\tilde{\mathbf{Y}}_{t+1} + \bar{\mathbf{x}}, \Gamma_t + \bar{\gamma}, \omega_t) - h(\mathbf{Y}_t + \bar{\mathbf{x}}, \Gamma_t + \bar{\gamma}, \omega_t)) \quad .\end{aligned}$$

$\Phi_{t+1} = (\mathbf{Y}_{t+1}, \Gamma_{t+1}, \omega_{t+1})$ is a function of only $\mathbf{Y}_t$, Γ_t, ω_t, and i.i.d. vectors $\mathbf{U}_{t+1}$. Therefore, $(\Phi_t)_{t\in\mathbb{N}}$ is a homogeneous Markov chain.

A.4 Proof of Corollary 1

By definition, we have

$$\begin{aligned}h(\mathbf{x}_{\text{opt}} + \alpha\mathbf{x}, \gamma_{\text{opt}} + \alpha\gamma, \omega) &= \underbrace{f(\mathbf{x}_{\text{opt}} + \alpha\mathbf{x})}_{A} \\ &+ \underbrace{\sum_{i=1}^m (\gamma_{\text{opt}}^i + \alpha\gamma^i) g_i(\mathbf{x}_{\text{opt}} + \alpha\mathbf{x})}_{B} + \underbrace{\sum_{i=1}^m \frac{\omega^i}{2} g_i(\mathbf{x}_{\text{opt}} + \alpha\mathbf{x})^2}_{C} \quad .\end{aligned}$$

By developing A, B, and C, we obtain

$$\begin{aligned}A &= \alpha^2 f(\mathbf{x}_{\text{opt}} + \mathbf{x}) + (1 - \alpha^2) f(\mathbf{x}_{\text{opt}}) + \alpha(1 - \alpha) \underbrace{\mathbf{x}_{\text{opt}}^\intercal \mathbf{H}}_{\nabla_{\mathbf{x}} f(\mathbf{x}_{\text{opt}})} \mathbf{x} \quad , \\ B &= \sum_{i=1}^m \alpha^2 (\gamma_{\text{opt}}^i + \gamma^i) g_i(\mathbf{x}_{\text{opt}} + \mathbf{x}) + \alpha(1 - \alpha)\gamma_{\text{opt}}^i \underbrace{\mathbf{b}^i}_{\nabla_{\mathbf{x}} g_i(\mathbf{x}_{\text{opt}})} \mathbf{x} \quad , \\ C &= \alpha^2 \sum_{i=1}^m \frac{\omega^i}{2} g_i(\mathbf{x}_{\text{opt}} + \mathbf{x})^2 \quad .\end{aligned}$$

The constraints being active at $\mathbf{x}_{\text{opt}}$, $h(\mathbf{x}_{\text{opt}}, \gamma_{\text{opt}}, \omega) = f(\mathbf{x}_{\text{opt}})$ for all $\omega \in (\mathbb{R}_>^+)^m$. It follows that

$$\begin{aligned}&\mathcal{D}h_{\mathbf{x}_{\text{opt}},\gamma_{\text{opt}},\omega}(\mathbf{x}_{\text{opt}} + \alpha\mathbf{x}, \gamma_{\text{opt}} + \alpha\gamma) \\ &= \alpha^2 \Big(f(\mathbf{x}_{\text{opt}} + \mathbf{x}) + \sum_{i=1}^m (\gamma_{\text{opt}}^i + \gamma^i) g_i(\mathbf{x}_{\text{opt}} + \mathbf{x}) + \frac{\omega^i}{2} g_i(\mathbf{x}_{\text{opt}} + \mathbf{x})^2 \\ &- f(\mathbf{x}_{\text{opt}}) \Big) + \alpha(1 - \alpha) \underbrace{\Big(\nabla_{\mathbf{x}} f(\mathbf{x}_{\text{opt}}) + \sum_{i=1}^m \nabla_{\mathbf{x}} g_i(\mathbf{x}_{\text{opt}}) \Big)}_{\mathbf{0}} \mathbf{x} \quad .\end{aligned}$$

The KKT stationarity condition in (29) is satisfied for $\mathbf{x}_{\text{opt}}$ and γ_{opt}. Therefore,

$$\begin{aligned}&\mathcal{D}h_{\mathbf{x}_{\text{opt}},\gamma_{\text{opt}},\omega}(\mathbf{x}_{\text{opt}} + \alpha\mathbf{x}, \gamma_{\text{opt}} + \alpha\gamma) \\ &\qquad = \alpha^2 \mathcal{D}h_{\mathbf{x}_{\text{opt}},\gamma_{\text{opt}},\omega}(\mathbf{x}_{\text{opt}} + \mathbf{x}, \gamma_{\text{opt}} + \gamma) \quad .\end{aligned}$$

Consequently, $(\Phi_t)_{t\in\mathbb{N}}$ is a homogeneous Markov chain with f convex quadratic.

A.5 Proof of Theorem 3

We express $\frac{1}{t} \ln \frac{\|\mathbf{X}_t - \mathbf{x}_{\text{opt}}\|}{\|\mathbf{X}_0 - \mathbf{x}_{\text{opt}}\|}$, $\frac{1}{t} \ln \frac{\|\gamma_t - \gamma_{\text{opt}}\|}{\|\gamma_0 - \gamma_{\text{opt}}\|}$, and $\frac{1}{t} \ln \frac{\sigma_t}{\sigma_0}$ as a function of the homogeneous Markov chain $(\Phi_t)_{t\in\mathbb{N}}$ defined in Theorem 1. Using the property of the logarithm, we have

$$\begin{aligned}\frac{1}{t} \ln \frac{\|\mathbf{X}_t - \mathbf{x}_{\text{opt}}\|}{\|\mathbf{X}_0 - \mathbf{x}_{\text{opt}}\|} &= \frac{1}{t} \sum_{k=0}^{t-1} \ln \frac{\|\mathbf{X}_{k+1} - \mathbf{x}_{\text{opt}}\|}{\|\mathbf{X}_k - \mathbf{x}_{\text{opt}}\|} \\ &= \frac{1}{t} \sum_{k=0}^{t-1} \ln \frac{\|\mathbf{Y}_{k+1}\|}{\|\mathbf{Y}_k\|} \mathcal{G}_\sigma(1, \varsigma * \mathbf{U}_{k+1}) \\ &= \frac{1}{t} \sum_{k=0}^{t-1} \ln \|\mathbf{Y}_{k+1}\| - \frac{1}{t} \sum_{k=0}^{t-1} \ln \|\mathbf{Y}_k\| \\ &+ \frac{1}{t} \sum_{k=0}^{t-1} \ln \mathcal{G}_\sigma(1, \varsigma * \mathbf{U}_{t+1}) \quad . \qquad (39)\end{aligned}$$

$(\Phi)_{t\in\mathbb{N}}$ is positive Harris-recurrent with an invariant probability measure π and $E_\pi(|\ln \|[\phi]_1\||) < \infty$, $E_\pi(|\ln \|[\phi]_2\||) < \infty$, and $E_\pi(\mathcal{R}(\phi)) < \infty$. Therefore, we can apply Theorem 2 to the right-

hand side of (39). We obtain

$$\lim_{t\to\infty}\frac{1}{t}\ln\frac{\|\mathbf{X}_t-\mathbf{x}_{\text{opt}}\|}{\|\mathbf{X}_0-\mathbf{x}_{\text{opt}}\|}=\lim_{t\to\infty}\frac{1}{t}\sum_{k=0}^{t-1}\ln\|\mathbf{Y}_{k+1}\|$$
$$-\lim_{t\to\infty}\frac{1}{t}\sum_{k=0}^{t-1}\ln\|\mathbf{Y}_k\|+\lim_{t\to\infty}\frac{1}{t}\sum_{k=0}^{t-1}\ln\mathcal{G}_\sigma(1,\varsigma*\mathbf{U}_{t+1})$$
$$=\int\ln\|[\phi]_1\|\pi(d\phi)-\int\ln\|[\phi]_1\|\pi(d\phi)+\int\mathcal{R}(\phi)\pi(d\phi)$$
$$=-\text{CR}\ .$$

We proceed similarly with $\frac{1}{t}\ln\frac{\|\gamma_t-\gamma_{\text{opt}}\|}{\|\gamma_0-\gamma_{\text{opt}}\|}$ and $\frac{1}{t}\ln\frac{\sigma_t}{\sigma_0}$.

$$\frac{1}{t}\ln\frac{\|\gamma_t-\gamma_{\text{opt}}\|}{\|\gamma_0-\gamma_{\text{opt}}\|}=\frac{1}{t}\sum_{k=0}^{t-1}\ln\|\Gamma_{k+1}\|-\frac{1}{t}\sum_{k=0}^{t-1}\ln\|\Gamma_k\|$$
$$+\frac{1}{t}\sum_{k=0}^{t-1}\ln\mathcal{G}_\sigma(1,\varsigma*\mathbf{U}_{t+1})\ , \tag{40}$$

$$\frac{1}{t}\ln\frac{\sigma_t}{\sigma_0}=\frac{1}{t}\sum_{k=0}^{t-1}\frac{\sigma_{k+1}}{\sigma_k}$$
$$=\lim_{t\to\infty}\frac{1}{t}\sum_{k=0}^{t-1}\ln\mathcal{G}_\sigma(1,\varsigma*\mathbf{U}_{t+1})\ . \tag{41}$$

By applying Theorem 2 to the right-hand side of (40) and (41), we obtain

$$\lim_{t\to\infty}\frac{1}{t}\ln\frac{\|\gamma_t-\gamma_{\text{opt}}\|}{\|\gamma_0-\gamma_{\text{opt}}\|}=\lim_{t\to\infty}\frac{1}{t}\ln\frac{\sigma_t}{\sigma_0}=-\text{CR}\ .$$